Automated Discovery of Important Chemical Reactions

by

Colin A. Grambow

B.Eng. Chemical Engineering
McGill University (2015)

M.S. Chemical Engineering Practice
Massachusetts Institute of Technology (2017)

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . .
Department of Chemical Engineering
April 24, 2020

Certified by . . . .
William H. Green
Hoyt C. Hottel Professor in Chemical Engineering
Thesis Supervisor

Accepted by . . . .
Patrick S. Doyle
Robert T. Haslam Professor of Chemical Engineering
Singapore Research Professor
Chairman, Committee for Graduate Students

Automated Discovery of Important Chemical Reactions

by

Colin A. Grambow

Submitted to the Department of Chemical Engineering
on April 24, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Innovations in chemistry are often informed by decades of accumulated chemical knowledge encoded into manually constructed reaction templates and rules of reactivity. Examples include retrosynthetic analysis for organic synthesis planning; chemical reaction mechanism generation for complex combustion, pyrolysis, and low-temperature oxidation processes; and elucidation of low-energy catalytic pathways. Nonetheless, all known chemistry is dwarfed by the vastness of chemical space, most of which still lies unexplored. *De novo* reaction discovery is rare but presents an enormous potential to uncover novel synthetic routes and key pathways in reaction mechanisms. Automated potential energy surface exploration has become a promising method to search for new reaction pathways, albeit at the expense of costly quantum mechanical calculations.

Therefore, this thesis develops methods to enable more computationally efficient discovery while also correctly determining thermochemistry and kinetics to allow for the construction of accurate reaction mechanisms.

By utilizing automated transition state finding algorithms based on quantum chemistry, the thesis assesses which algorithm is most viable for the efficient discovery of new reactions, and it identifies key pathways of an important ketohydroperoxide system. It demonstrates that quantum chemical data can be used with emerging machine learning methods to estimate molecular thermochemistry. Leveraging a large data set of low-quality data in combination with a small data set of high-accuracy data in a transfer learning approach enables predictions that significantly improve upon group additivity methods, which are common in automated mechanism generation, and upon machine learning models that only use density functional theory data. Furthermore, an automated workflow is developed to further enhance high-level quantum chemistry calculations using bond additivity corrections.

While quantum chemistry calculations are incredibly useful at providing highly accurate data, their high cost—especially when applied to thousands of reaction pathways—limits their utility for discovering new chemistry. Therefore, this thesis improves the throughput of automated discovery via a combination of quantum chemistry data generation and reactivity prediction using deep learning. It automatically generates a data set of tens of thousands of elementary chemical reactions that are used to train a novel activation energy prediction model, which can quickly assess the importance of new reactions.

Thesis Supervisor: William H. Green
Title: Hoyt C. Hottel Professor in Chemical Engineering

Acknowledgments

Over the years, many people have been instrumental to my success at MIT, and I would certainly be remiss if I did not thank everyone that has had a part in it.

First and foremost, I am immensely grateful to my advisor, Bill Green, whose ability to inspire and drive research while single-handedly managing our large group has never ceased to amaze me. Besides fostering my independence and professional growth, his perennial positivity and encouragement for collaboration have made working in his group a joy.

Heather Kulik, Bernhardt Trout, and Klavs Jensen have been very supportive committee members. I greatly appreciated their guidance and assistance over the years. I especially want to thank Gwen Wilcox for her straightforward yet invariably kind attitude. Her logistical prowess made the many administrative tasks a breeze, and her caring for the students was greatly appreciated.

I have had the pleasure of working with many great collaborators throughout my PhD, both within and outside of my research group. On my first project, I worked together with Yury Suleimanov, Adeel Jamal, Yi-Pei Li, and Judit Zádor. I was constantly consulting with Yury, Adeel taught me much of what I now know about quantum chemistry, and Yi-Pei was a great resource both on that project and for all of my subsequent work. My machine learning adventures got started because of the inspiring work Kehang Han was doing. I have worked together closely with Lucky Pattanaik on all of my machine learning projects, as our excessively long Slack history can attest to. I enjoyed working with Mark Payne, with whom I also spent many hours fixing servers and discussing programming. Max Liu and Matt Johnson were especially helpful for any questions about RMG. Alon Dana, Duminda Ranasinghe, Phalgun Lolur, and Yanfei Guan assisted in many different ways, from teaching me more about quantum chemistry to providing me with new data. To all the other Green Group members not mentioned here by name: Know that I have greatly appreciated interacting with all of you and am grateful that you have made the group such a pleasant community to be a part of.

I am very thankful for my great first-year class, who even managed to make the difficult first semester enjoyable. I am especially grateful to Sam, Max, Kim, Alan, Falco, Will, and Christina who have remained good friends ever since. I also want to thank everyone in my practice school group and on GSC-X my year.

None of the past five years would have been possible without the constant love and support from my parents, Julie and Kai, and my brother, Brendan. I greatly cherish the enthusiasm and encouragement they have shown, and continue to show, for every decision I have made in my life.

And last, but certainly not least, to my partner, Ki-Joo: Thank you for supporting me through all the ups and downs. I cannot imagine the last four years without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Innovations in chemistry are often informed by decades of accumulated chemical knowledge encoded into manually constructed reaction templates and rules of reactivity. An example is the computer-generation of chemical kinetic models in complex systems, such as combustion and pyrolysis, the importance of which has long been recognized in the scientific community.[1–3] Detailed (microkinetic) mechanisms attempt to enumerate all relevant elementary reaction steps along with their rates. By modeling a system in this manner, the potential for extrapolation to new conditions and systems is greatest and individual reactions can be investigated with quantum chemistry. Nonetheless, all known chemistry is dwarfed by the vastness of chemical space, most of which still lies unexplored.[4] *De novo* reaction discovery is rare but presents an enormous potential to uncover novel, key pathways in reaction mechanisms.

Intricate chemical processes may require mechanisms with thousands of species and more than ten thousand reactions.[5,6] Figure 1.1 illustrates the sizes of some published mechanisms and how they have grown over the years. As a result of the large model sizes, which require thousands of parameters, manual construction of detailed chemical mechanisms is very tedious and often not tractable.[7] Automated reaction mechanism generation[8–11] very significantly reduces the amount of manual effort required and has enabled successful generation of complex radical-driven mechanisms for many different systems.[12–21]

Automated reaction mechanism generation is only possible if thousands of thermochemical and kinetic parameters can be rapidly estimated to sufficient accuracy. While several estimation algorithms exist,[11,22–24] limited data often do not afford the required accuracy and more sophisticated algorithms would be needed. Automated reaction mechanism generation also relies on reaction templates that are based on rules of reactivity manually established over decades of research.[11] Therefore, it is not suitable for discovering novel chemistry, especially for systems that have not had the same amount of attention as combustion and pyrolysis. For example, new reactions that were discovered by Magoon et al.[25] significantly altered the results reported by Vandewiele et al.[16] in their study of jet fuel pyrolysis, and the novel pathway reported by Jalan et al.[26] has been shown

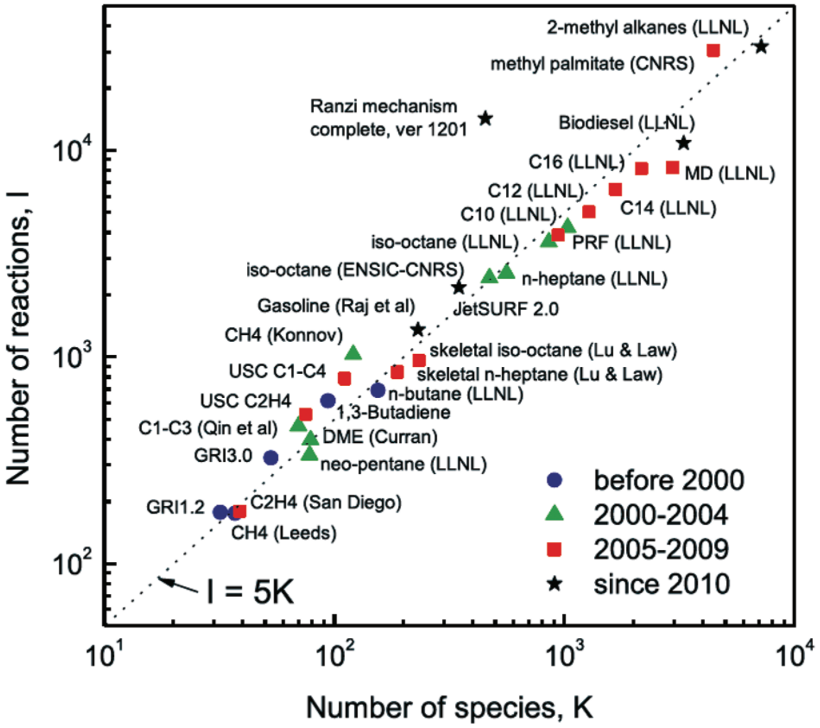


Figure 1.1. Typical sizes of published kinetic models.[5,6] Reprinted by permission from Springer Nature: *Science China Chemistry* "Challenges and Perspectives of Combustion Chemistry Research" Yuan W., Li Y., Qi F., © 2017.

to be influential in low-temperature oxidation.[27] To enable routine discovery of new reactions across many systems, methods for exploring unknown chemical space—without human bias—become necessary. The goal of this thesis is to provide methods for finding unknown reactions automatically and to enable more accurate estimation of parameters relevant in automated mechanism generation.

1.1 Automated reaction mechanism generation

Simply stated, a reaction mechanism is a list of reactions, ideally elementary, along with their kinetic parameters and thermochemical parameters for all of the species involved in the reactions. Several software packages exist for generating reaction mechanisms automatically. They include *Reaction*,[8] EXGAS,[9] Genesys,[10] and the Reaction Mechanism Generator (RMG).[11]

RMG uses tens of reaction templates, also known as families, in a rate-based algorithm[28] to generate a chemical mechanism.[11] Figure 1.2 illustrates a simplified version of the basic RMG expansion algorithm. The initial species with their respective concentrations are loaded into the "core". In each iteration, all possible reactions between the species in the core are generated using the reaction families, which may lead to new species that populate the "edge". The system is then simulated as an isothermal batch reactor and the edge species with the highest production rate is added to the core at the end of the iteration. This procedure is repeated until certain termination

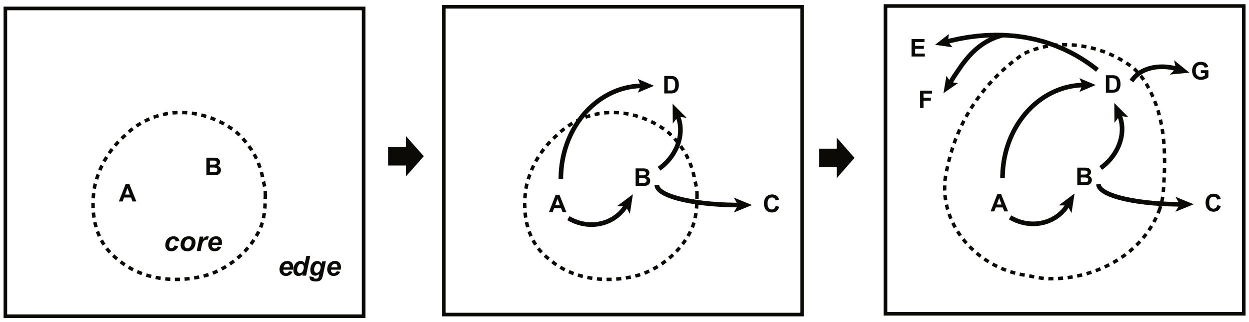


Figure 1.2. Expansion of an RMG model.[11] In each iteration, all possible reactions are generated between species in the core and the edge species with the highest flux is moved to the core. The figure is licensed under CC BY 4.0, © 2016 Gao C.W., Allen J.W., Green W.H., West R.H.

criteria are satisfied. To obtain the reaction rates necessary for simulating the system in each iteration, RMG computes the rate constant of a reaction by averaging across a hierarchical tree of rate estimation rules.[11] A separate tree exists for each reaction family. As reactions are generally assumed to be reversible, the reverse rate constant is calculated in a thermodynamically consistent manner using[29]

$$\frac{k_\mathrm{f}}{k_\mathrm{r}} = K_\mathrm{eq}^\circ = \exp\left[-\frac{\Delta_\mathrm{rxn}G^\circ(T)}{RT}\right] \tag{1.1}$$

k_f is the forward rate constant, k_r is the reverse rate constant, K_eq° is the equilibrium constant, $\Delta_\mathrm{rxn}G^\circ(T)$ is the Gibbs free energy change of reaction, R is the gas constant, and T is the temperature. $\Delta_\mathrm{rxn}G^\circ(T)$ can be computed using the thermochemical parameters of each species: enthalpy of formation, entropy, and heat capacity. RMG estimates these parameters using group additivity[22,23] and the hydrogen bond increment method.[24]

As already alluded to previously, the existing RMG reaction templates may not encode all necessary transformations, thereby precluding the incorporation of novel reactions. Furthermore, parameter estimation errors during the mechanism growth phase may cause important reactions to not be included if the production rate of a species is underestimated. Therefore, a methodology for discovering novel chemistry and improved parameter/property estimation are desirable.

1.2 Automated discovery of chemical reactions

In order to discover novel chemical reactions that are not restricted to the heuristics imposed by software like RMG, experimental or first-principles (*ab initio*) methods become necessary. However, some kinetically significant reactions might be very difficult to detect and measure experimentally because of very low concentrations or difficulties in isolating specific reactions.[7] Theoretical methods, including quantum chemistry and molecular dynamics, naturally isolate reactions and concentration dependence is not an issue.

In quantum chemistry, *ab initio* discovery is possible by solving the time-independent Schrödinger

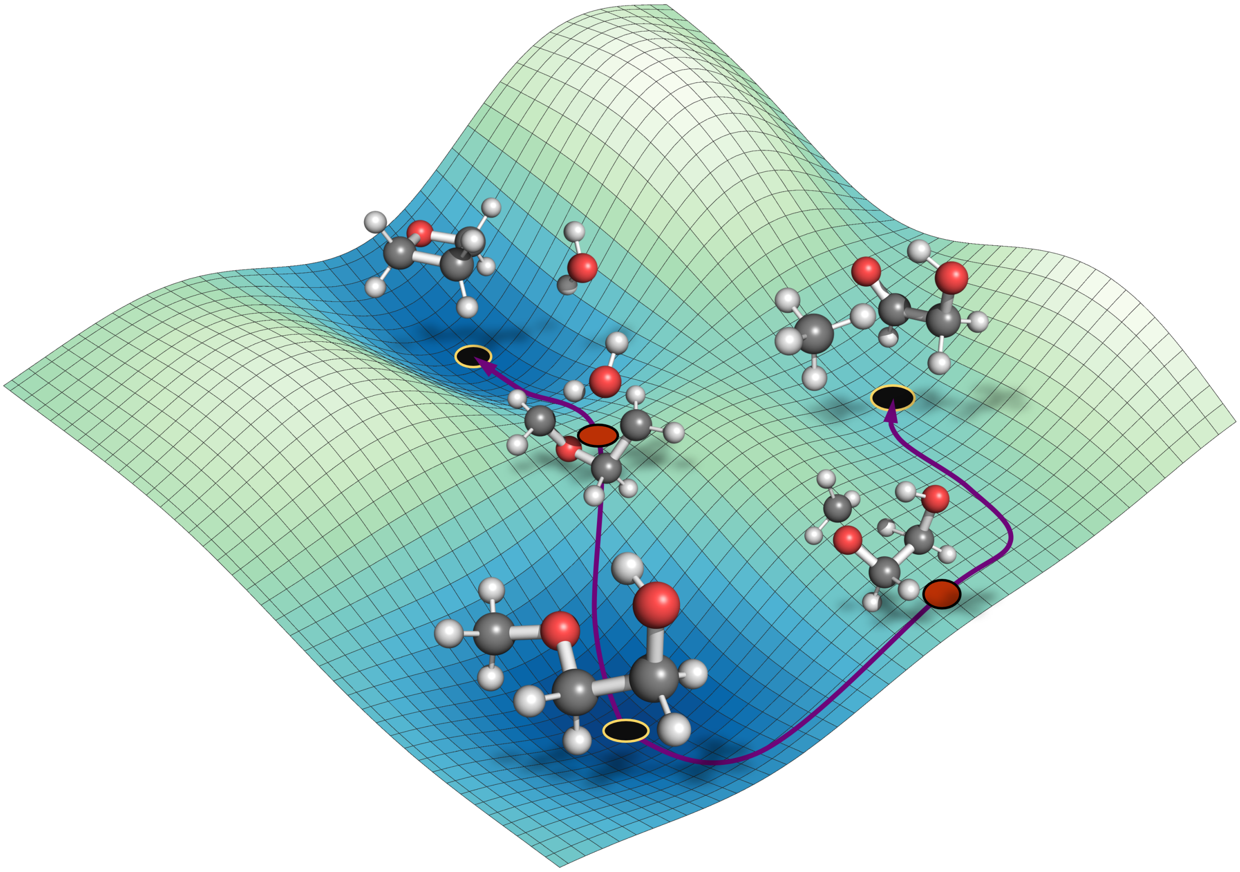


Figure 1.3. Simplified representation of a reactive potential energy surface. The black dots mark the location of minima corresponding to reactant and products, and the red dots mark the location of saddle points corresponding to transition states. The arrows follow the minimum energy paths.

equation for several nuclear configurations, which produces a potential energy surface (PES). The equation is given by[30]

$$\hat{H}\Psi = E\Psi \tag{1.2}$$

$\hat{H}$ is the Hamiltonian operator, Ψ is the wave function, and E is the electronic energy. In general, Equation (1.2) cannot be solved analytically and the PES is of very high dimensionality, which leads to large computational cost. As a result, many different approximation methods exist, with popular ones ranging from density functional theory (DFT) to high-accuracy coupled cluster theory. A simplified representation of a potential energy surface in two dimensions is shown in Figure 1.3. The most important points on the surface are stationary points, where the force acting on atoms is zero. A minimum, as indicated by the valleys in Figure 1.3, corresponds to a stable chemical species and a first-order saddle point, which is a maximum in one direction and a minimum in all other directions, corresponds to a transition state. The steepest descent path going in both directions from the saddle point is the reaction path known as the minimum energy path (MEP). Computing the second derivative matrix (Hessian) at stationary points yields the information necessary for computing statistical mechanical partition functions. The partition function of a stable species gives rise to its thermochemistry[29] and the partition functions of reactant and transition state enable calculating the rate constant for the reaction using transition state theory (TST).[31–33]

Minimizations to obtain optimized intermediates (reactants and products) can be completed routinely and systematically,[34] but obtaining transition states is associated with significantly more difficulty due to the non-convex optimizations to first-order saddle points. In general, the problem is to find all relevant transition states and associated products given a reactant. The high dimensionality of the PES greatly complicates this search. For use in automated reaction mechanism generation, only the reactions with large enough rates are required, which most often correspond to the reactions with small potential energy barrier heights. Many different transition state search algorithms exist, of which a non-exhaustive set is described in Chapter 2. Many additional methods, several based on molecular dynamics,[35–38] are being actively developed. Recent reviews provide an overview of many of the methods.[39–41]

1.3 Machine learning in chemistry

Machine learning has undoubtedly had a tremendous impact in recent years with many examples making headlines outside of the scientific community. Popular reinforcement learning models have revolutionized many games: AlphaZero has achieved performances in chess and Go that were previously not thought possible.[42] The follow-up model, MuZero, can achieve better-than-grandmaster performance even when the rules of the game are not provided.[43] Even complex video games like StarCraft II are starting to be dominated by machine learning models.[44] In other areas, the BERT natural language processing model,[45] human face generation,[46] and medical radiograph diagnosis[47] have also made the news. A significant contributor to the machine learning boom was the seminal performance by Krizhevsky et al. on the ImageNet classification task.[48]

Machine learning in chemistry has not yet enjoyed the same level of popularity as it has in computer science but Figure 1.4 shows that this is rapidly changing. Machine learning is a very broad term for many different parameter estimation techniques, which include quantitative structure-activity relationships (QSAR) that have been popular in chemistry for a long time.[49] Deep learning generally refers to the use of deep artificial neural networks, which have the benefit that they do not require manual feature engineering, which is required for traditional machine learning methods. They also strongly benefit from the large data sets that are becoming more and more prevalent in the chemical and biological sciences. Their training time only scales linearly with the number of training data and the time to evaluate a deep learning model is independent of the training data size. While often considered to be a "black box" model, deep neural networks can be integrated with physically meaningful representations in chemistry (e.g., graph convolutions[50–52]) that are starting to enable physical interpretation.

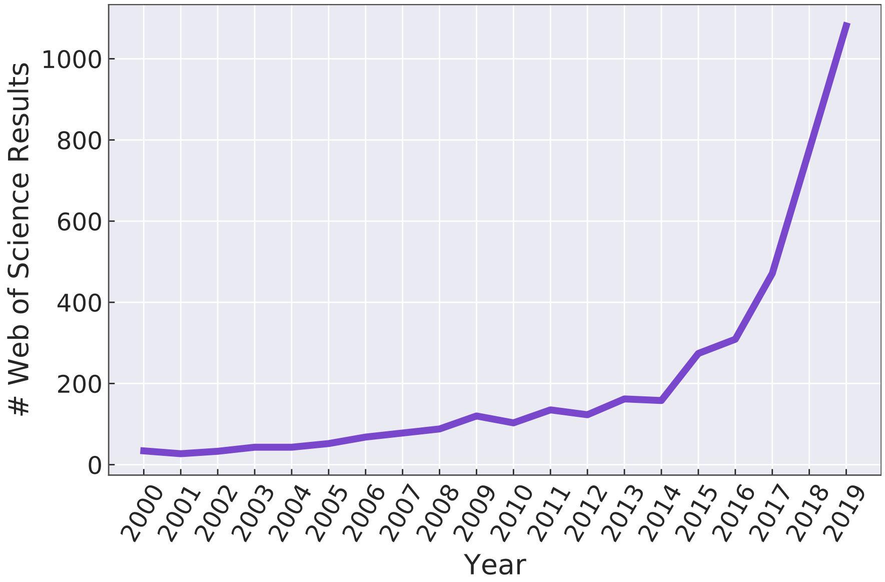


Figure 1.4. Number of Web of Science articles for *machine learning in chemistry* related topics from 2000 to 2019 (the exact search term was *("machine learning" OR "artificial intelligence" OR "deep learning") AND chemi**).

As evidenced by Figure 1.4, summarizing the entire literature on deep learning in chemistry is nearly impossible, but there exist many current reviews. They mostly deal with organic synthesis planning (retrosynthesis, reaction prediction, product ranking),[53–58] drug/materials design and discovery,[54–57,59–61] and quantum chemistry.[54,55,57–59,62] Less frequent topics include structural biology,[59] spectroscopy,[54] literature extraction,[54] sensors,[55] deep generative modeling to generate and optimize molecules,[63] optimization of reaction parameters and process conditions,[57,58] analytical chemistry and catalysis,[58] and applications in chemical engineering.[64]

Kinetic mechanism elucidation has so far been mostly limited to the methods described in Section 1.1, but deep learning methods are promising for enhancing automated mechanism generation by enabling more efficient discovery of novel reactions and improved kinetic and thermochemical parameter estimation. In fact, the poor scaling of electronic structure methods with molecular size mentioned in Section 1.2 may preclude purely quantum-chemistry-based discovery, whereas new deep learning techniques coupled with the increasing availability of chemical data sets may greatly improve the scalability of reaction discovery. However, quantitative data for high-accuracy molecular property prediction and reaction discovery is generally still lacking.

1.4 Thesis overview

My thesis has made several contributions toward developing methods to uncover novel reactivity and to enable improved estimation of thermochemical parameters. Chapter 2 examines reaction discovery using only quantum chemistry. Chapter 3 introduces machine learning methods to create a thermochemistry estimator using molecular graph input. Chapter 4 builds upon Chapter 2 by generating a large data set of reactions using quantum chemistry and Chapter 5 uses this data to learn a model of chemical reactivity. The following outlines the chapters in more detail.

Chapter 2 investigates automated transition state finding algorithms and assesses their viability for *ab initio* automated reaction discovery. It focuses on a ketohydroperoxide important in liquid phase autoxidation and in gas phase partial oxidation and pre-ignition chemistry for which only a few pathways are known due to its low concentration, instability, and various analytical chemistry limitations. We discovered 75 elementary-step unimolecular reactions through a combination of DFT with several automated transition state search algorithms: the Berny algorithm coupled with the freezing string method, single- and double-ended growing string methods, KinBot, and the single-component artificial force induced reaction method. This joint approach significantly outperformed previous manual and automated transition state searches—68 of the discovered reactions were previously unknown and completely unexpected. We showed that the low-barrier chemical reactions involve promising new chemistry that may be relevant in atmospheric and combustion systems. The chapter highlights the complexity of chemical space exploration and the advantage of combined application of several approaches. When a combined approach is not feasible, the single most promising method was the single-ended growing string method.

Chapter 3 develops a new thermochemistry estimator that leverages small, but high-quality data sets to provide accurate estimates for automated reaction mechanism generation. For deep neural networks to be effective, they require large training data sets which are only available at low levels of theory, whereas automated reaction mechanism generation frequently requires data at the level of high-accuracy quantum chemistry methods beyond DFT. To overcome these limitations, we calculated new high-level data sets and derived bond additivity corrections to significantly improve enthalpies of formation. We adopted a transfer learning technique to train neural network models for the prediction of thermochemical parameters that achieve good performance even with relatively small sets of high-accuracy data. The training data for the entropy model was carefully selected so that important conformational effects were captured. Along with the newly developed thermochemistry predictors, we implemented an automated method for deriving bond additivity corrections that can be used to enhance high-quality data generation in the future.

Chapter 4 uses the most promising automated transition state finding method in Chapter 2, the single-ended growing string method, to generate a data set of tens of thousands of elementary chemical reactions. It thereby addresses the current scarcity of quantitative chemical reaction data, especially of atom-mapped reactions. We used automated potential energy surface exploration to generate 12 000 organic reactions involving H, C, N, and O atoms calculated at the ωB97X-D3/def2-TZVP quantum chemistry level by performing geometry optimizations and frequency calculations for reactants, products, and transition states of all reactions. Additionally, we extracted atom-mapped reaction SMILES, activation energies, and enthalpies of reaction. We showed that the data is made up of very diverse reactions spanning across a wide range of activation energies, including many reactions of kinetic relevance.

Chapter 5 addresses the problem of the large computational cost of quantum chemistry calculations for finding unknown reactions. Using the data developed in Chapter 4, we developed a deep learning model to train an activation energy predictor which can quickly assess the importance of new candidate reactions. We constructed the model in a chemically meaningful way such that it predominantly learns from the parts of the molecules that contribute most to the activation energy. Moreover, we constructed the model in a way such that it is not restricted to specific reaction templates as mentioned in Section 1.1, but can produce an estimate for any atom-mapped reaction.

Chapter 6 highlights some of the limitations of the research in this thesis and the current state-of-the-art and accordingly proposes future directions of study.

1.5 References

(1) Tomlin, A. S.; Turányi, T.; Pilling, M. J., Mathematical tools for the construction, investigation and reduction of combustion mechanisms In *Low-temperature Combustion and Autoignition, Volume 35*, Pilling, M. J., Ed., 1st edition; Elsevier: Amsterdam, 1997.

(2) Battin-Leclerc, F. Detailed Chemical Kinetic Models for the Low-Temperature Combustion of Hydrocarbons with Application to Gasoline and Diesel Fuel Surrogates. *Prog. Energy Combust. Sci.* **2008**, *34*, 440–498.

(3) Battin-Leclerc, F.; Blurock, E.; Bounaceur, R.; Fournet, R.; Glaude, P. A.; Herbinet, O.; Sirjean, B.; Warth, V. Towards Cleaner Combustion Engines Through Groundbreaking Detailed Chemical Kinetic Models. *Chem. Soc. Rev.* **2011**, *40*, 4762–4782.

(4) Hoffmann, T.; Gastreich, M. The Next Level in Chemical Space Navigation: Going Far Beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**, *24*, 1148–1156.

(5) Lu, T.; Law, C. K. Toward Accommodating Realistic Fuel Chemistry in Large-Scale Computations. *Prog. Energy Combust. Sci.* **2009**, *35*, 192–215.

(6) Yuan, W.; Li, Y.; Qi, F. Challenges and Perspectives of Combustion Chemistry Research. *Sci. China Chem.* **2017**, *60*, 1391–1401.

(7) Broadbelt, L. J.; Pfaendtner, J. Lexicography of Kinetic Modeling of Complex Reaction Networks. *AIChE J.* **2005**, *51*, 2112–2121.

(8) Blurock, E. S. Reaction: System for Modeling Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 607–616.

(9) Warth, V.; Battin-Leclerc, F.; Fournet, R.; Glaude, P. A.; Côme, G. M.; Scacchi, G. Computer Based Generation of Reaction Mechanisms for Gas-Phase Oxidation. *Comput. Chem.* **2000**, *24*, 541–560.

(10) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M. F.; Marin, G. B. Genesys: Kinetic model construction using chemo-informatics. *Chem. Eng. J.* **2012**, *207-208*, 526–538.

(11) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(12) Allen, J. W.; Scheer, A. M.; Gao, C. W.; Merchant, S. S.; Vasu, S. S.; Welz, O.; Savee, J. D.; Osborn, D. L.; Lee, C.; Vranckx, S.; Wang, Z.; Qi, F.; Fernandes, R. X.; Green, W. H.; Hadi, M. Z.; Taatjes, C. A. A Coordinated Investigation of the Combustion Chemistry of Diisopropyl Ketone, a Prototype for Biofuels Produced by Endophytic Fungi. *Combust. Flame* **2014**, *161*, 711–724.

(13) Prozument, K.; Suleimanov, Y. V.; Buesser, B.; Oldham, J. M.; Green, W. H.; Suits, A. G.; Field, R. W. A Signature of Roaming Dynamics in the Thermal Decomposition of Ethyl Nitrite: Chirped-Pulse Rotational Spectroscopy and Kinetic Modeling. *J. Phys. Chem. Lett.* **2014**, *5*, 3641–3648.

(14) Carr, A. G.; Class, C. A.; Lai, L.; Kida, Y.; Monrose, T.; Green, W. H. Supercritical Water Treatment of Crude Oil and Hexylbenzene: An Experimental and Mechanistic Study on Alkylbenzene Decomposition. *Energy Fuels* **2015**, *29*, 5290–5302.

(15) Gao, C. W.; Vandeputte, A. G.; Yee, N. W.; Green, W. H.; Bonomi, R. E.; Magoon, G. R.; Wong, H. W.; Oluwole, O. O.; Lewis, D. K.; Vandewiele, N. M.; Van Geem, K. M. JP-10 Combustion Studied with Shock Tube Experiments and Modeled with Automatic Reaction Mechanism Generation. *Combust. Flame* **2015**, *162*, 3115–3129.

(16) Vandewiele, N. M.; Magoon, G. R.; Van Geem, K. M.; Reyniers, M. F.; Green, W. H.; Marin, G. B. Kinetic Modeling of Jet Propellant-10 Pyrolysis. *Energy Fuels* **2015**, *29*, 413–427.

(17) Class, C. A.; Liu, M.; Vandeputte, A. G.; Green, W. H. Automatic Mechanism Generation for Pyrolysis of Di-*tert*-butyl Sulfide. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21651–21658.

(18) Seyedzadeh Khanshan, F.; West, R. H. Developing Detailed Kinetic Models of Syngas Production from Bio-Oil Gasification Using Reaction Mechanism Generator (RMG). *Fuel* **2016**, *163*, 25–33.

(19) Wagner, A. L.; Yelvington, P. E.; Cai, J.; Green, W. H. Combustion of Synthetic Jet Fuel: Chemical Kinetic Modeling and Uncertainty Analysis. *J. Propuls. Power* **2017**, *33*, 350–359.

(20) Vervust, A. J.; Djokic, M. R.; Merchant, S. S.; Carstensen, H. H.; Long, A. E.; Marin, G. B.; Green, W. H.; Van Geem, K. M. Detailed Experimental and Kinetic Modeling Study of Cyclopentadiene Pyrolysis in the Presence of Ethene. *Energy Fuels* **2018**, *32*, 3920–3934.

(21) Chu, T. C.; Buras, Z. J.; Oßwald, P.; Liu, M.; Goldman, M. J.; Green, W. H. Modeling of Aromatics Formation in Fuel-Rich Methane Oxy-Combustion with an Automatically Generated Pressure-Dependent Mechanism. *Phys. Chem. Chem. Phys.* **2019**, *21*, 813–832.

(22) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *Chem. Rev.* **1958**, *29*, 546–572.

(23) Benson, S. W., *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*, 2nd edition; Wiley: New York, 1976.

(24) Lay, T. H.; Bozzelli, J. W.; Dean, A. M.; Ritter, E. R. Hydrogen Atom Bond Increments for Calculation of Thermodynamic Properties of Hydrocarbon Radical Species. *J. Phys. Chem.* **1995**, *99*, 14514–14527.

(25) Magoon, G. R.; Aguilera-Iparraguirre, J.; Green, W. H.; Lutz, J. J.; Piecuch, P.; Wong, H.-W.; Oluwole, O. O. Detailed Chemical Kinetic Modeling of JP-10 (*exo*-Tetrahydrodicyclopentadiene) High-Temperature Oxidation: Exploring the Role of Biradical Species in Initial Decomposition Steps. *Int. J. Chem. Kinet.* **2012**, *44*, 179–193.

(26) Jalan, A.; Alecu, I. M.; Meana-Pañeda, R.; Aguilera-Iparraguirre, J.; Yang, K. R.; Merchant, S. S.; Truhlar, D. G.; Green, W. H. New Pathways for Formation of Acids and Carbonyl Products in Low-Temperature Oxidation: The Korcek Decomposition of γ-Ketohydroperoxides. *J. Am. Chem. Soc.* **2013**, *135*, 11100–11114.

(27) Ranzi, E.; Cavallotti, C.; Cuoci, A.; Frassoldati, A.; Pelucchi, M.; Faravelli, T. New Reaction Classes in the Kinetic Modeling of Low Temperature Oxidation of n-Alkanes. *Combust. Flame* **2015**, *162*, 1679–1691.

(28) Susnow, R. G.; Dean, A. M.; Green, W. H.; Peczak, P.; Broadbelt, L. J. Rate-Based Construction of Kinetic Models for Complex Systems. *J. Phys. Chem. A* **1997**, *101*, 3731–3740.

(29) Ruscic, B.; Bross, D. H., Thermochemistry In *Mathematical Modelling of Gas-Phase Complex Reaction Systems: Pyrolysis and Combustion*, Faravelli, T., Manenti, F., Ranzi, E., Eds.; Computer-Aided Chemical Engineering, Vol. 45; Elsevier: Cambridge, MA, 2019.

(30) McDouall, J. J. W., *Computational Quantum Chemistry: Molecular Structure and Properties in Silico*; The Royal Society of Chemistry: Cambridge, UK, 2013.

(31) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

(32) Evans, M. G.; Polanyi, M. Some Applications of the Transition State Method to the Calculation of Reaction Velocities, Especially in Solution. *Trans. Faraday Soc.* **1935**, *31*, 875.

(33) Wigner, E. The Transition State Method. *Trans. Faraday Soc.* **1938**, *34*, 29–41.

(34) Bera, P. P.; Sattelmeyer, K. W.; Saunders, M.; Schaefer III, H. F.; Schleyer, P. R. Mindless Chemistry. *J. Phys. Chem. A* **2006**, *110*, 4287–4290.

(35) Vázquez, S. A.; Otero, X. L.; Martinez-Nunez, E. A Trajectory-Based Method to Explore Reactions. *Molecules* **2018**, *23*, 3156.

(36) Döntgen, M.; Schmalz, F.; Kopp, W. A.; Kröger, L. C.; Leonhard, K. Automated Chemical Kinetic Modeling via Hybrid Reactive Molecular Dynamics and Quantum Chemistry Simulations. *J. Chem. Inf. Model.* **2018**, *58*, 1343–1355.

(37) Fu, C. D.; Pfaendtner, J. Lifting the Curse of Dimensionality on Enhanced Sampling of Reaction Networks with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2018**, *14*, 2516–2525.

(38) Jara-Toro, R. A.; Pino, G. A.; Glowacki, D. R.; Shannon, R. J.; Martínez-Núñez, E. Enhancing Automated Reaction Discovery with Boxed Molecular Dynamics in Energy Space. *ChemSystemsChem* **2020**, *2*, e190024.

(39) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123*, 385–399.

(40) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **2020**, *71*, 121–142.

(41) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *8*, e1354.

(42) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play. *Science* **2018**, *362*, 1140–1144.

(43) Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T.; Silver, D. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. **2019**, arXiv: 1911.08265.

(44) Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.; Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gulcehre, C.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; Silver, D. Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* **2019**, *575*, 350–354.

(45) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. **2018**, arXiv: 1810.04805.

(46) Karras, T.; Laine, S.; Aila, T., A Style-Based Generator Architecture for Generative Adversarial Networks In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, 2019, pp 4396–4405.

(47) Rajpurkar, P.; Irvin, J.; Ball, R. L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C. P.; Patel, B. N.; Yeom, K. W.; Shpanskaya, K.; Blankenberg, F. G.; Seekins, J.; Amrhein, T. J.; Mong, D. A.; Halabi, S. S.; Zucker, E. J.; Ng, A. Y.; Lungren, M. P. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLoS Med.* **2018**, *15*, e1002686.

(48) Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks In *Advances in Neural Information Processing Systems 25*, Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc.: 2012, pp 1097–1105.

(49) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'Min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(50) Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. **2015**, arXiv: 1509.09292.

(51) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

(52) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**, arXiv: 1704.01212.

(53) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(54) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

(55) Rodrigues Jr, J. F.; Florea, L.; de Oliveira, M. C. F.; Diamond, D.; Oliveira Jr, O. N. A Survey on Big Data and Machine Learning for Chemistry. **2019**, arXiv: 1904.10370.

(56) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.

(57) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem. Int. Ed.*, DOI: 10.1002/anie.201909987.

(58) Cova, T. F.; Pais, A. A. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* **2019**, *7*, 809.

(59) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307.

(60) Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to Machine Learning: Recent Approaches to Materials Science–A Review. *J. Phys. Mater.* **2019**, *2*, 032001.

(61) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128.

(62) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347.

(63) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—A Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849.

(64) Venkatasubramanian, V. The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally? *AIChE J.* **2019**, *65*, 466–478.

Chapter 2

Unimolecular reaction pathways of a γ-ketohydroperoxide from combined application of automated reaction discovery methods

This work has previously appeared as Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zádor, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048. The artificial force induced reaction method calculations were enabled by Adeel Jamal and the KinBot calculations were done by Judit Zádor. Yi-Pei Li wrote the code to align reactant and product structures. Yury Suleimanov analyzed many of the results. The code for reproducing the results is available at `https://github.com/cgrambow/AutomaticReactionDiscovery`.

2.1 Introduction

For several decades, chemists have been trying to develop theoretical approaches to predict the reactivity of specific chemical compounds and to guide chemical discovery. Recent advances in electronic structure theory and high-performance computer technologies should make it possible to achieve this long-standing goal, and so achieve a much better understanding of systems where multiple reactions are occurring simultaneously.[1,2] Complex chemistry is common in combustion chemistry,[3] polymerization,[4] catalysis,[5–7] and environmental processes,[8] all of which are subject to continuous study due to their fundamental and industrial importance. A fundamental understanding of all chemical compounds and elementary reactions of a given chemical process can facilitate the design of more effective technologies.[9] In general, multiple competing reaction paths exist, which lead to a variety of products, especially if initial species are highly reactive (e.g., radicals, peroxides, and catalytic intermediates) or if a system is at high temperature. Historically, new reactions were discovered experimentally by the serendipitous detection of unexpected products. New reactions can also be discovered serendipitously on the computer.[10–12] Recently, Zimmerman has intentionally discovered several important new reactions using quantum chemistry.[13,14] However, existing search techniques are CPU-intensive and sole reliance on experiments to find new reactions is insufficient. With modern computational capabilities, it should become possible to discover all the important reaction pathways more reliably and more rapidly than is possible using experiments alone.

A potential energy surface (PES) is a multidimensional function of atomic coordinates that provides comprehensive information on all reaction paths. Its local minima correspond to reactants, intermediates, and products. These are generally easy to characterize due to simple chemical bonding rules (in most cases) making it relatively easy to predict their 3D geometries. Numerical optimization of these geometries is straightforward because the negative of the gradient along the PES always points downhill (i.e., it is a local minimization problem).[15] In fact, automated searches for local and global minima have already proved to be very successful.[16] A more challenging task is to detect and characterize first-order saddle points that connect local minima along minimum energy paths (MEPs), which are necessary to describe most transition states (TSs), the key regions of the PES to calculate reaction rate coefficients. Predicting geometries of TSs is more difficult. Numerical procedures for saddle points must step uphill in one certain direction (the reaction coordinate, usually unknown *a priori*) and downhill in all other orthogonal directions—a more challenging task than finding a local minimum.[17] Typically, many low-energy saddle points exist in the vicinity of minima, which correspond to torsional rearrangements and lead to different conformations of a given minimum structure. Reactive TSs are mostly higher in energy and correspond to a change in bonding. Additionally, if one saddle point can be found that directly connects two structures in a reactive event, often many additional saddle points exist, several of them representing conformers of the TS.[18] Another concern is that the dimension of a PES ($3N - 6$) increases with the number of atoms (N) in the system. Therefore, construction and subsequent global mapping of PESs have prohibitive computational costs even for reactive systems consisting of only 5–6 atoms. At present, most saddle points are found by human-guided exploration of reaction pathways. Human expectations and chemical intuition bias such an approach, usually limiting the search to expected reaction paths. The process is also slow and tedious. It is therefore highly desirable to develop efficient methods for automatically searching for unexpected TSs and corresponding reaction pathways on PESs.

Several effective methods for automatically searching reaction pathways with given reactant(s) have been proposed recently. Some of these methods are based on adding an external temperature/pressure control or artificial forces in the initial reactive system, which drive the reaction to occur in the direction of different products. These include, but are not limited to, metadynamics,[19] an *ab initio* nanoreactor,[20] and the artificial-force-induced reaction (AFIR) method.[21] While the former two are based on molecular dynamics (MD) with special techniques to accelerate the evolution of the reactive system, the artificial force induced reaction (AFIR) strategy of the global reaction route mapping (GRRM) method utilizes specially designed minimization functions that are composed of the adiabatic PES and an artificial force term. In the case of a unimolecular initial reactant channel, this strategy, called single-component AFIR (SC-AFIR), utilizes local optimization procedures between two fragments to explore the reactions possible due to intramolecular pathways. Several methods which are conceptually similar to AFIR have also been proposed. These are the single-ended growing string method (SSM),[22] in which several nodes along coordinates (defined in terms of bonds, angles, and torsions) are added to the reactant(s) to drive the search towards a desired product, and the coordinate driving method (CD),[23] which uses constrained electronic structure optimizations along a series of proposed reaction coordinates in order to detect feasible reaction pathways.

An alternative strategy is to use a two-step approach. During the first step, a set of possible product channels is generated using graphical (or combinatorial) rules based on the concept of the chemical bond[17,24] or using a heuristic generation of high-energy reactive complexes followed by their relaxation to minima.[25] During the second step, the algorithm attempts to connect them with the reactant channel using double-ended saddle point search methods, such as freezing string (FSM)[17,26] or growing string (GSM) methods.[24,27] Nudged elastic band methods could also be used.[28,29] It is also possible to heuristically generate guess structures for the TSs[30] which are further refined by conventional methods, such as the Berny method,[31–33] as is implemented in the KinBot program.[34,35] If one TS has been identified, programs such as MSTor can automatically search for its conformers,[36,37] and the rate constant can be calculated from the set of conformers, e.g., by multistructural transition state theory.[38]

In addition, it is worth noting that some methods based on machine learning algorithms have recently been proposed for the prediction of organic chemistry reactions.[39] However, the predictive capabilities of such approaches are implicitly limited by the range of reaction types contained in the training set used, and usually more explicitly limited by the use of specific reaction templates. Therefore, these methods are unlikely to discover a new and unexpected types of reactions.

While the aforementioned methods have been successfully applied to some organic (and organometallic) reactions for searching reaction pathways,[2,3,13,14,17,40] it is difficult to assess which methods are most effective, since the simulations were performed separately for different systems using different algorithms. There is a paucity of comparative studies and understanding of reaction discovery algorithm performance. In the present work, we aim to address this issue by performing a joint study of the unimolecular decomposition and isomerization of a γ-ketohydroperoxide (KHP), due to the importance of this class of molecules in autoxidation and low-temperature combustion chemistry.[10,41–43] We have already studied the chemistry of this KHP using the Berny method and FSM for which six unimolecular decomposition saddle points were found.[17] Thus, this system also serves as a reference point for the methods that were available to us when we initiated this project and that we selected for the present calculations. This includes one single-ended (SSM) and two double-ended (FSM and GSM) methods, the heuristic KinBot algorithm, and SC-AFIR. In the following sections, we discuss the rich chemistry discovered using these methods and the advantages of such a joint approach. We also compare the performance and analyze the main pitfalls of each method. We hope that the results of this work will lead to the development of more cost-effective and reliable automated reaction discovery methods for general application in complex chemical systems.

2.2 Computational procedures

2.2.1 Combinatorial search using string methods and berny optimization

The computational procedure for the automated identification of reaction pathways using string methods is nearly identical to the one previously proposed by us for the freezing string method (FSM).[17] In the present study, we considered only breaking and forming a maximum of three bonds and only allowed products with permissible valences (e.g., maximum of four for carbon). In addition, the breaking of double (or higher order) bonds was not permitted. Products were only added to the set of new products if their connectivity was different from all other products and if they were not isomorphic with the reactant molecule (i.e., conformational changes were not counted as reactions). This procedure resulted in a set of 4324 possible product structures.

Reaction Mechanism Generator (RMG) thermodynamic libraries[44] and Benson's group additivity approach[45] were used to estimate the standard reaction enthalpies, $\Delta H_\mathrm{r}^\circ$, for all the generated reactions. We did not attempt to discover reactions with an estimated $\Delta H_\mathrm{r}^\circ$ higher than $20\,\mathrm{kcal\,mol^{-1}}$. Following this filtering step, 562 product structures remained. In order to verify whether the group additivity estimates were a sufficient proxy for the true reaction energies, more accurate estimates using density functional theory (DFT) were calculated for the set of filtered reactions. This analysis showed that group additivity was sufficient for the reaction filtering step, although several reactions were included in the filtered set that would have been excluded based on the DFT criterion, and it is possible that some additional reactions would have been included; however, we chose to only consider the set of 562 products, as this already constitutes a very broad search space.

Initial geometries of the reactant and product(s) were generated by constructing linear connections based on rules corresponding to the hybridization of atoms and by constructing ring structures from templates as implemented in the Open Babel program.[46] The energies of generated structures were minimized using the MMFF94 force field also implemented in the program. When the product channels contained two or three fragments, they were translated and rotated in space in order to maximize the overlap with the initial reactant molecule while constrained by a minimum distance from each other. The dihedral angles of rotatable bonds were also modified to further maximize the overlap.

The reactant and product conformers were optimized using Berny optimization at the B3LYP/6-31+G* level of theory implemented in the Gaussian 09 program.[47] Searches for guess transition state (TS) structures were initiated using the three string methods: FSM, GSM, and SSM.

For FSM, we used the same default spacing parameters as in our previous study[17] (20 nodes, 6 perpendicular gradients per node). For this work, we rewrote the FSM algorithm, included more termination criteria checks and interfaced it with Gaussian 09. We also added one restarting option which we found to provide more stability to the FSM calculations: When the electronic structure calculation failed, the previous gradient was used for a given node. This was not done if the node had just been generated in the interpolation step of the FSM algorithm where a previous gradient calculation was not yet available. We also set the maximum number of nodes to 40 in order to terminate the FSM calculation when it becomes stuck and ends up oscillating back and forth between geometries.

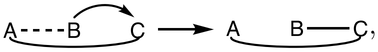
For GSM and SSM, we also used standard spacing parameters and settings (100 maximum iterations, exactly 11 nodes for GSM, a maximum of 30 nodes for SSM with a minimum node spacing of 1.0 Å) and used the program developed by one of the authors of those methods.[48] After successful generation of the string paths, the first reactive peak was selected as an initial guess TS structure provided that at least one bond changed in the structure, which was determined using automatic identification of bonds as implemented in Open Babel.[46]

The next step was the optimization of these guess structures using the Berny optimization algorithm as implemented in Gaussian 09. While the GSM and SSM code already incorporates a Hessian-free exact saddle point search, we nevertheless reoptimized the structures with the Berny algorithm for consistency across all methods. This was required anyway for GSM paths where the first reactive barrier closest to the reactant was not the maximum barrier in the entire GSM path, since the GSM code only completes an exact saddle point search for the highest energy structure. Finally, intrinsic reaction path (IRC) calculations were performed to verify whether the detected saddle point corresponded to the expected reactant and product paths.

2.2.2 Single-component artificial force induced reaction method

For SC-AFIR,[21,49] we used the M06-2X/3-21G level of theory[50] to enable inexpensive searches along the SC-AFIR paths. To be consistent with the other methods, we then reoptimized all critical points from the SC-AFIR search with B3LYP/6-31+G* using Berny optimization in Gaussian 09, including IRC calculations leading to products. It should be noted that only mechanistic studies were considered and gradient calls are not reported here for the SC-AFIR calculations. We set the γ-value of the SC-AFIR function to 400 kJ mol^{-1} to encompass a large search space, and considered all artificial forces to every fragment within KHP through intramolecular interactions.

2.2.3 Heuristic KinBot approach

KinBot is a heuristic search program, which proposes sensible guesses for certain types of very broad reaction classes based on hard-won chemical knowledge.[34,35] For instance, during internal abstractions of atoms or groups, symbolically 

, the atom that is being placed from one connectivity to another (B) is typically halfway between the original A and final C atoms, with further prescriptions for bond angles, for instance depending on the relative positions of A and C in the molecule, or the type of the atoms involved. Creating these prescribed centers automatically and systematically essentially obviates human effort and eliminates human mistakes (hence the name, KinBot = "Kinetics roBot"), while still capitalizing on the knowledge we gathered for certain types of reactions. Note that KinBot was initially created and optimized for reactions involving C, H, and O atom containing radicals. Creating heuristics for radicals is a much simpler task than creating ones for closed shell molecules, because the sensible, i.e., low-energy, pathways in a radical decomposition or isomerization almost always involve the radical center. For closed shell molecules, such as the KHP in this study, it is harder to predict (and code) the preferred pathways.

In the KinBot case, we analyzed the synergetic effect of the joint application of several approaches. For this purpose, the corresponding calculations were performed in two steps. During the first run, similar to the other methods, we used KinBot's default reaction types and parameters without any prior knowledge about the KHP chemistry. In the second run, we extended the reaction types and included some variations to allow for more proposed structures than previously while taking into account the chemical reactions detected in the first run and by the other four methods (FSM, GSM, SSM, and SC-AFIR). Most significantly, in the first run, KinBot's constraints limited the search space primarily to the transfer of H atoms to other atoms, but in the second run KinBot considered transferring every atom type to every other one since many reactions of this type have been observed with the other methods. As a result of relaxing KinBot's criteria, the possibilities were significantly expanded resulting in the detection of significantly more channels. Because of the different nature of these calculations, we will distinguish the KinBot results from the first calculations (as a "blind method assessment") and the second calculations (as an "extended guided run") throughout the chapter as well as in the relevant figures and table. Note that a similar "refining" approach could be implemented for the other methods. For instance, in the case of the string methods (FSM/GSM/SSM), inclusion of zwitterionic structures in the initial set of product channels as well as inclusion of channels with $\Delta H_\mathrm{r}^\circ$ higher than $20\,\mathrm{kcal\,mol^{-1}}$ would lead to the detection of more channels. However, in the present study, we limit ourselves only to illustrative refining of KinBot calculations due to the flexibility of the KinBot algorithm and the ease of implementing modifications in the corresponding code.

In Figure 2.1 we summarize the broad reaction types invoked by KinBot for KHP. It is important to note, however, that many times the intended reactions do not happen, but often the calculations

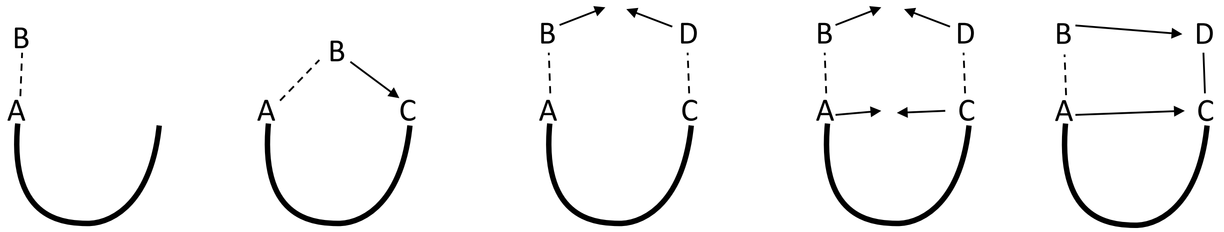


Figure 2.1. The schematics of the reactions invoked by KinBot automatically for KHP. The exact bond lengths, angles, and dihedrals depend on the type of the atoms in the active center and the length of the chain between A and C.

converge rapidly to a slightly different, yet chemically significant saddle point. It is possible to increase the number of templates further to allow for even more complex rearrangements, but we did not extend beyond the ones in Figure 2.1.

When constructing the corresponding 3D structures for the reaction types in Figure 2.1, moving the atoms in the reactive center to their desired positions all at once is not feasible in most cases, because it causes the other, spectator atoms to clash, which, in turn leads to the rapid fragmentation or chaotic rearrangement of the structure in the subsequent optimization steps. To achieve these generic, but prescribed steric structures in the reactive center of the molecule, KinBot manipulates some large-amplitude motions, typically involving rotations around the appropriate dihedral angles, in a stepwise fashion. In each step, a constrained optimization is carried out at a very cheap level of theory to lead the structure approximately along a MEP-like valley. We used HF/STO-3G (first, generic run) and AM1 (second, expanded run) as cheap levels in this work, both of which have negligible computational costs, especially considering that in these constrained optimizations (all bond lengths are frozen) only a greatly reduced dimensionality gradient is needed for most steps. Once the desired conformation was achieved, we invoked the Berny algorithm to optimize the structure to a first-order saddle point at the B3LYP/6-31G* level of theory. Finally, for the successfully optimized structures we used IRC calculations to identify the reactant and the products, considering success if the reactant is the initial structure and the product is not. KinBot does not check whether the found saddle point is the intended one or not. KinBot also uses Gaussian 09[47] to carry out both the constrained and the final optimizations.

2.3 Comparative analysis of methods

Although the comparative analysis of the methods presented in this section focuses on rather technical aspects of the present calculations, we note that automated reaction discovery methodology is in an early stage and such an analysis can significantly improve the capability of discovering new and important chemistry.

2.3.1 Computational/statistical details

The statistical details of the automated search calculations are summarized in Table 2.1. This includes the total number of channels tested, total number of gradient calls (excluding IRC), total number of detected valid (i.e., reactive) saddle points, number of channels with wrong reactants, number of conformational saddle points, number of crashed searches, and the sources of failures. The statistics on the energy barriers and detected chemical reactions are discussed in the following subsections. Table 2.1 shows that SSM detected the highest number of chemical reactions corresponding to elementary steps of converting γ-ketohydroperoxide; 371 out of 562 channels resulted in detecting valid (i.e., reactive) saddle points. However, most of these channels correspond to duplicates (either exact duplicates or duplicates due to the same reaction with different but equivalent hydrogens) and only 50 saddle points are truly unique. Nevertheless, SSM provides the highest number of unique saddle points among all methods tested in the present work with GSM discovering 46 unique saddle points. However, Table 2.1 shows that the cost for such success is rather high. The computational expenses of these two growing string methods, estimated as the total number of gradient calls, were more than an order of magnitude higher than those of the FSM calculations. From this perspective, FSM represents an inexpensive alternative; 39 unique saddle points were detected and the total number of gradient calls was smaller by a factor of ten compared to GSM and SSM.

Rough estimates using all geometries of the detected saddle points in the KinBot study show that the Hessian matrix calculations are approximately five times more expensive than the corresponding gradient calculations at the high level of theory (B3LYP/6-31+G*), therefore, the initial construction of the Hessian adds only a negligible cost to its updates, which require only the gradients. The cost of the highly constrained Hessian calculations used to pre-optimize the structures is negligible at the HF/STO-3G (first run) or AM1 (second run) levels. This means that the computational expense of KinBot (first run) was about eight times less than that of FSM, with the total number of saddle points detected by KinBot being slightly lower (32 in the first run). When KinBot was run with the extended set of rules, it was still about three times cheaper than FSM and detected more saddle points (48). Unfortunately, we were unable to extract the total number of gradient calls for the SC-AFIR calculations as we did not have access to the corresponding source code. The total number of unique saddle points detected by SC-AFIR was rather small (7).

For the three string methods, the ratio of the average number of gradient calls per successfully detected reaction is 308 (FSM):793 (GSM):1221 (SSM). These numbers take into account both the string method and Berny optimization steps. This ratio is in line with previous estimates for string method calculations,[22,24,26,27] which indirectly shows that all three string methods provide equally good saddle point guess structures for subsequent Berny optimization in cases where the saddle point was eventually successfully detected. However, it is clearly different from the ratio for the

Table 2.1. Summary of the automated detection of elementary chemical reaction steps using string methods (FSM, GSM, SSM), SC-AFIR, and KinBot.

Method	**FSM**			**GSM**			**SSM**			**SC-AFIR**			**KinBot** First run (Generic)			**KinBot** Second run (Expanded)		
Channels (N_{channel})	562			562			562			74[l]			200			443		
Gradients (w/o IRC)	65 420[a]			756 227[a]			649 589[a]			N/A			8046[a] (10 093[b])			23 458[a] (25 333[c])		
Valid TSs (N_{detected})	153	27%[d]	–	215	38%[d]	–	371	66%[d]	–	17	23%[d]	–	83	42%[d]	–	162	37%[d]	–
• Unique	39	7%[d]	26%[e]	46	8%[d]	21%[e]	50	9%[d]	14%[e]	7	10%[d]	41%[e]	32	16%[d]	39%[e]	48	11%[d]	30%[e]
• Duplicates[g]	102	18%[d]	67%[e]	155	28%[d]	72%[e]	297	53%[d]	80%[e]	10	14%[d]	59%[e]	51	26%[d]	61%[e]	114	26%[d]	70%[e]
• Equivalent hydrogens[h]	12	2%[d]	8%[e]	14	2%[d]	6%[e]	24	4%[d]	6%[e]	0	0%[d]	0%[e]	0	0%[d]	0%[e]	0	0%[d]	0%[e]
• Intended[i]	19	3%[d]	12%[e]	24	4%[d]	11%[e]	27	5%[d]	7%[e]	14	19%[d]	82%[e]	–	–	–	–	–	–
• Unintended	134	24%[d]	88%[e]	191	34%[d]	89%[e]	344	61%[d]	93%[e]	3	4%[d]	18%[e]	–	–	–	–	–	–
Wrong reactant	85	15%[d]	–	39	7%[d]	–	23	4%[d]	–	4	5%[d]	–	13	6%[d]	–	22	5%[d]	–
Conformational	3	0.5%[d]	–	13	2%[d]	–	0	0%[d]	–	53	72%[d]	–	48	24%[d]	–	69	16%[d]	–
Crashed (N_{crashes})	321	57%[d]	–	295	52%[d]	–	168	30%[d]	–	–	–	–	56[m]	28%[d]	–	190[m]	43%[d]	–
• Berny failure	230	41%[d]	72%[f]	36	6%[d]	12%[f]	20	4%[d]	12%[f]	–	–	–	–	–	–	–	–	–
• Gradient failure[j]	91	16%[d]	28%[f]	88	16%[d]	30%[f]	124	22%[d]	74%[f]	–	–	–	–	–	–	–	–	–
• Maximum iterations	0	0%[d]	0%[f]	171	30%[d]	58%[f]	18	3%[d]	11%[f]	–	–	–	–	–	–	–	–	–
• Dissociative[k]	0	0%[d]	0%[f]	6	1%[d]	2%[f]	6	1%[d]	4%[f]	–	–	–	–	–	–	–	–	–

[a] B3LYP/6-31+G*. [b] Gradients in HF/STO-3G constrained minimization steps. [c] Gradients in AM1 constrained minimization steps. [d] $1/N_{\text{channel}} \times 100\%$ (percentage of total number of channels). [e] $1/N_{\text{detected}} \times 100\%$ (percentage of total valid saddle points). [f] $1/N_{\text{crashes}} \times 100\%$ (percentage of total crashed searches). [g] Product and saddle point have the same adjacency matrix as a unique reaction product and saddle point. [h] Reaction is a duplicate because energetically equivalent hydrogens undergo the same reaction but with different atom indices. [i] For double-ended FSM and GSM: reaction that connects reactant and product structure specified in the input; for single-ended SSM: reaction that contains the bond changes that correspond to the driving coordinates specified in the input; for SC-AFIR: reaction where the IRC calculation at the high level matches the products predicted by the low-level AFIR calculation; KinBot does not test this. [j] Umbrella term for any error in the electronic structure program during a single gradient calculation. [k] Combinatorially generated reactions do not include barrierless dissociation reactions so these channels are placed together with crashed searches. [l] For SC-AFIR, the total number of channels is the number of first-step channels as given by the low-level M06-2X/3-21G search. [m] KinBot crashed searches includes internal coordinate failures in the pre-optimization phase and any type of error in the Berny optimization.

number of gradient calls in the corresponding total calculations from Table 2.1 suggesting that our procedure of detecting saddle points can be optimized in the future. For instance, Table 2.1 shows that many string method calculations resulted in crashed searches. FSM calculations crashed in more than 50% of cases; 321 out of 562 channels finished with an error. Interestingly, the source of errors is different for all string methods. For FSM, the crashed searches were primarily due to failure of the Berny optimization step (∼70%). Within the Berny optimization failures, almost 60% were due to reaching the maximum number of optimization steps (100), which signifies the inability to converge to a transition state structure in a reasonable amount of time. Approximately 20% of Berny optimization failures resulted from redundant internal coordinate errors in Gaussian 09. It is possible that such errors could be prevented by restarting the optimizations, but they typically are a result of chemically unreasonable geometries. Fewer than 20% of Berny optimization failures were due to convergence issues in the SCF calculations. Apparently, the maximum from the FSM path often provides a poor guess structure for further optimization[17] and often leads to crashes. Moreover, such guess structures can lead to saddle points which do not connect directly to the initial reactant; 85 wrong saddle points were detected in the FSM calculations. Note that GSM and SSM also detected a high number of saddle points for wrong reactant channels, 39 and 23, respectively. In such cases, GSM string paths exhibit multiple elementary steps with apparently bad guess structures for further Berny optimization.

For GSM, a high number of crashes (295) was also observed, although most of them originated from reaching the maximum number of iterations inside the GSM algorithm (100 per channel), most probably due to the impossibility of constructing a growing string path in a limited number of iterations in cases where the string contains multiple elementary steps. It should be noted that such a path profile does not imply that a single-elementary-step profile does not exist for the same reaction, but GSM does not attempt to exclusively search for single-step pathways in its current implementation. Among the string methods, SSM exhibited the lowest amount of crashes, mainly from convergence failures encountered in the external quantum chemistry package Gaussian 09 during gradient calculations. While analyzing/fixing failures of external software packages is not included in the scope of the present study, the other sources of crashes can be considered as "essential" failures of the algorithms as they require many gradient calls and lead to undesired results. To summarize our comparison between the string methods, we point out that SSM provides the highest number of detected saddle points, although many of them are duplicates. It also exhibits the lowest number of crashes and undesired channels, that is, wrong reactant(s) or conformational saddle points. However, it requires more gradient calls per successfully detected channel than the other string methods.

For SC-AFIR, we located 74 channels resulting from KHP as a reactant, 53 of which lead back to the reactant. The IRC scans and analysis of atom indices show that all of these channels are torsional conformation changes, that is, they do not result from bond scission and rearrangements. The reason for such a result is likely due to the different nature of the SC-AFIR algorithm, which treats saddle points relatively equally, whereas the string methods more strongly bias the search space by directly specifying reactants and products to only obtain reactive saddle points. Similarly, KinBot automatically performs conformational rearrangements before directly searching for reactive saddle points, thus leading to a smaller percentage of conformational saddle points relative to the total number of channels. The level of theory used for the SC-AFIR calculations (M06-2X/3-21G), which may enable more accurate treatment of long-range dispersion interactions resulting from intramolecular hydrogen bonding interactions involving any of the three oxygen atoms, may also somewhat influence the SC-AFIR search,[51] although further study would be required to validate this hypothesis. It is possible that dispersion interactions may only become significant with a larger basis set. After Berny optimization of saddle point structures at the B3LYP/6-31+G* level, 72% of the channels found were conformational saddle points (see Table 2.1). Yet, from the 17 saddle points corresponding to chemical reactions, 41% were unique. Meanwhile, 59% were exact duplicates (which includes different TS conformers), suggesting that SC-AFIR was relatively efficient at locating unique saddle points after the high level optimization. Additionally, it is possible that some saddle points that exist on the B3LYP/6-31+G* PES may not exist on the M06-2X/3-21G PES. Recently, Maeda and Harabuchi showed that SC-AFIR settings can be tuned in order to discover many more reactions and result in a performance that is comparable to the other methods.[52]

For KinBot, the percentage of crashed searches is comparable to the string methods (28% and 43% in the first and second runs, respectively) but the number of undesired saddle points (conformational or leading to wrong reactant(s)) is higher (61 and 91 in the first and second runs, respectively), originating mainly from conformational saddle points (48 and 69 in the first and second runs, respectively), that is, points that do not correspond to a real chemical reaction event. KinBot is most efficient in that it has the highest percentage of detected unique saddle points relative to its total number of channels searched. KinBot crashed for two main reasons: (1) the created guess structure was too far from any saddle points, so the number of maximum iterations (100) was reached in the Berny optimization, or (2) the Cartesian geometry modifications prescribed by KinBot could not be completed in redundant coordinates in Gaussian. While the first mode of failure is clearly an intrinsic property of KinBot, that is, not all saddle points can be predicted using templates, the second mode of failure can be prevented by taking smaller steps between consecutive geometry modification steps or imposing a looser convergence criterion for the intermediate step. We applied such techniques to minimize the second failure mode in the second run. In addition, in the second set of KinBot calculations finer IRC profiles were constructed leading to the detection of new channels. For example, a finer IRC profile may result in the detection of a biradical structure, which otherwise would not be detected if the IRC calculation steps past the shallow biradical well. It is interesting to note that although the total number of channels for the extended KinBot run was significantly larger than the total number of channels in the first run, 443 versus 200 (see Table 2.1), there were a significantly larger number of crashes, thus only resulting in 162 additional valid channels while yielding 16 additional unique channels. Evidently, some of the additional channels that KinBot explored in the extended run were less chemically feasible and resulted in worse initial saddle point structures, which then ran into errors during the Berny optimization step. This behavior is more similar to that observed with the string methods in that it renders KinBot less efficient at finding valid channels compared to the first run, but it significantly increases the number of unique saddle points found.

2.3.2 Statistics on energy barriers

Figure 2.2a shows the histograms of the energy barriers of the reactions detected by FSM, GSM, SSM, SC-AFIR, and KinBot (first generic run). We are reporting energy differences relative to the

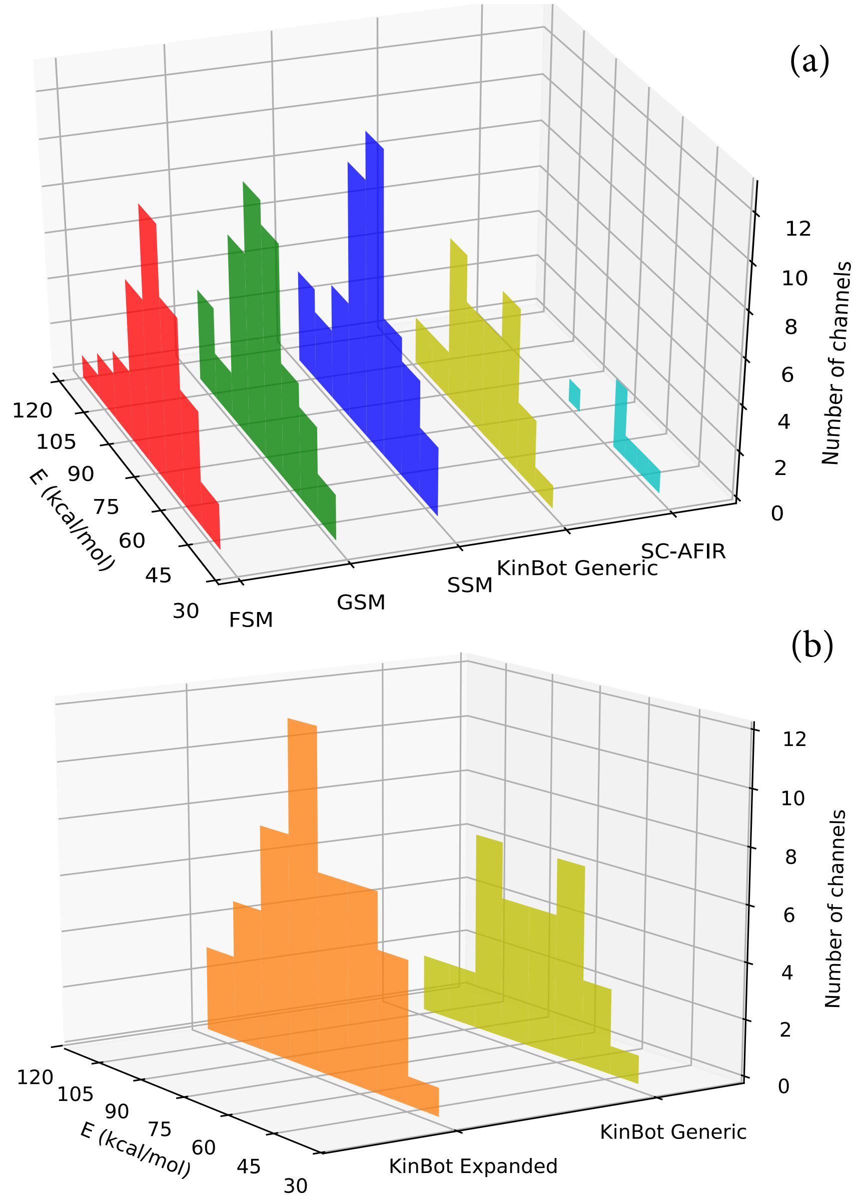


Figure 2.2. Histograms of the energy barriers (E_a) of the chemical reactions of unimolecular decomposition and isomerization of γ-ketohydroperoxide (with 8 equal-width bins containing the number of detected saddle points within each bin). (a) Reactions detected by FSM, GSM, SSM, SC-AFIR, and KinBot (first, generic run). (b) Reactions detected by KinBot during the first (generic) and second (expanded) runs.

reactant *excluding* the zero-point energy. Note that small differences in the energy barriers and reaction energies occur because the located transition states connect two relevant isomers along all different reaction pathways and may be different for each method. For FSM and GSM, the intended channel is the reaction that connects reactant and product structures specified in the input; for SSM it is the reaction that contains the bond changes that correspond to the driving coordinates specified in the input; for SC-AFIR it is the reaction where the IRC calculation at the high level matches the products predicted by the low-level AFIR calculation. However, the energy range of the conformers of the reactant was less than 5 kcal mol^{-1} for the set of reactant structures for all discovered reactions. All string methods have comparable energy barrier distributions and demonstrate a large peak in the intermediate range. The spread of energy barriers discovered by GSM is larger than that of FSM and KinBot, while SSM exhibits both the largest spread and finds the lowest barrier saddle points among the string methods. KinBot tends to avoid very high-barrier saddle points and is good at finding fairly low-barrier reactions. Figure 2.2b shows that this tendency remained in the extended run in which significantly more channels (48) were detected. However, as a result of extending the reaction rules such that all atom types are transferred to all other atom types, many more high-energy reactions were found in the second run, thus creating a histogram more similar to that of SSM. SC-AFIR also tends to find the lowest energy barriers; 6 out of 7 detected reactions have an energy barrier below 55 kcal mol^{-1}.

Figure 2.3 shows how the barrier heights obtained from the string methods compare to the barrier heights after Berny optimization of the saddle points. Clearly, FSM differs significantly from both GSM and SSM in that most of the FSM barriers are much greater than the corresponding optimized barrier heights. This again highlights the low quality of saddle point guess structures obtained from the FSM algorithm and explains the large number of crashes in the subsequent Berny optimization. Due to the exact saddle point search implemented in both growing string methods, the barrier heights of their guess geometries agree much better with the Berny optimized barrier heights for both GSM and SSM. Interestingly, there exist some GSM barriers that are lower in energy than the corresponding Berny barriers. This stems from the fact that an exact saddle point search is only initiated for the highest energy structure in the GSM code if the string is composed of multiple elementary steps. If the first elementary step in the string does not contain the maximum energy point, then the saddle point of the first step is not optimized and may lie in a lower region on the PES. Compared to FSM, the GSM saddle point guess structures are of much higher quality, as indicated by energies closer to the Berny optimized energies, even though some of them have not yet undergone the exact saddle point search as part of the string method (as just discussed). This is due to the gradual convergence of the string to a minimum energy path with subsequent GSM iterations, which does not take place in FSM, in which all structures are frozen in place after initial optimization. For SSM, only single step paths exist due to the nature of the algorithm, such that all SSM barriers are larger than the Berny optimized barriers (within numerical accuracy). For both GSM and SSM, some points lie relatively high above the line of equality. These are mostly points where the predicted saddle points from GSM/SSM converged to a chemically different saddle point

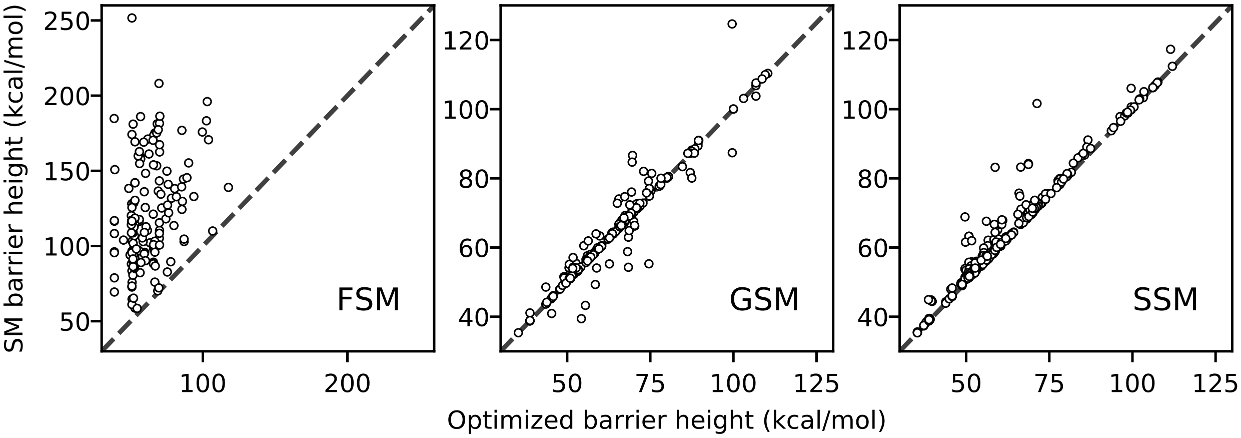


Figure 2.3. Comparison of the barrier heights from the three string methods (SM) to the optimized barrier heights after Berny optimization. The dashed line represents the line of equality.

after Berny optimization. This is likely a result of the differences in the saddle point optimization algorithm implemented in GSM and SSM versus the Berny algorithm. It is also possible that some error is incurred due to the fact that GSM and SSM only construct an approximate Hessian, whereas the first step in the Berny optimization is the calculation of an analytical Hessian.

2.4 Analysis of identified chemical reactions

The previously outlined procedure found 75 distinct transition states originating from 3-hydroperoxypropanal, 68 of which were completely unexpected. Each search algorithm except for SC-AFIR discovered several TSs not found by any of the other methods. Figure 2.4 summarizes all identified products of the unimolecular one-step decomposition and isomerization of KHP obtained by all five methods (with two sets of results for KinBot).

In total, 75 unique reactions were identified using all five methods. This significantly outperforms our previous FSM study[17] where only six chemical reactions were detected (reactions 4, 37, 47, 59, 69, and 72), which is partially due to the difficulties with the previous FSM calculations mentioned above as well as a result of the fact that the overwhelming majority of new chemical reactions have high-energy saddle points. In the previous study, such FSM profiles were discarded. Interestingly, in the present study, reactions 4 and 59 were detected by the other methods and not by FSM. Apparently, such discrepancy between the present and previous FSM calculations originates from the sensitivity of FSM to slight changes in starting geometries and the subsequently generated suboptimal saddle point guess structures. Another possible explanation is the different level of theory used previously (M06-2X/6-311++G*).[17] In addition, the reverse channel of reaction 17 was published in literature this work was being conducted.[53]

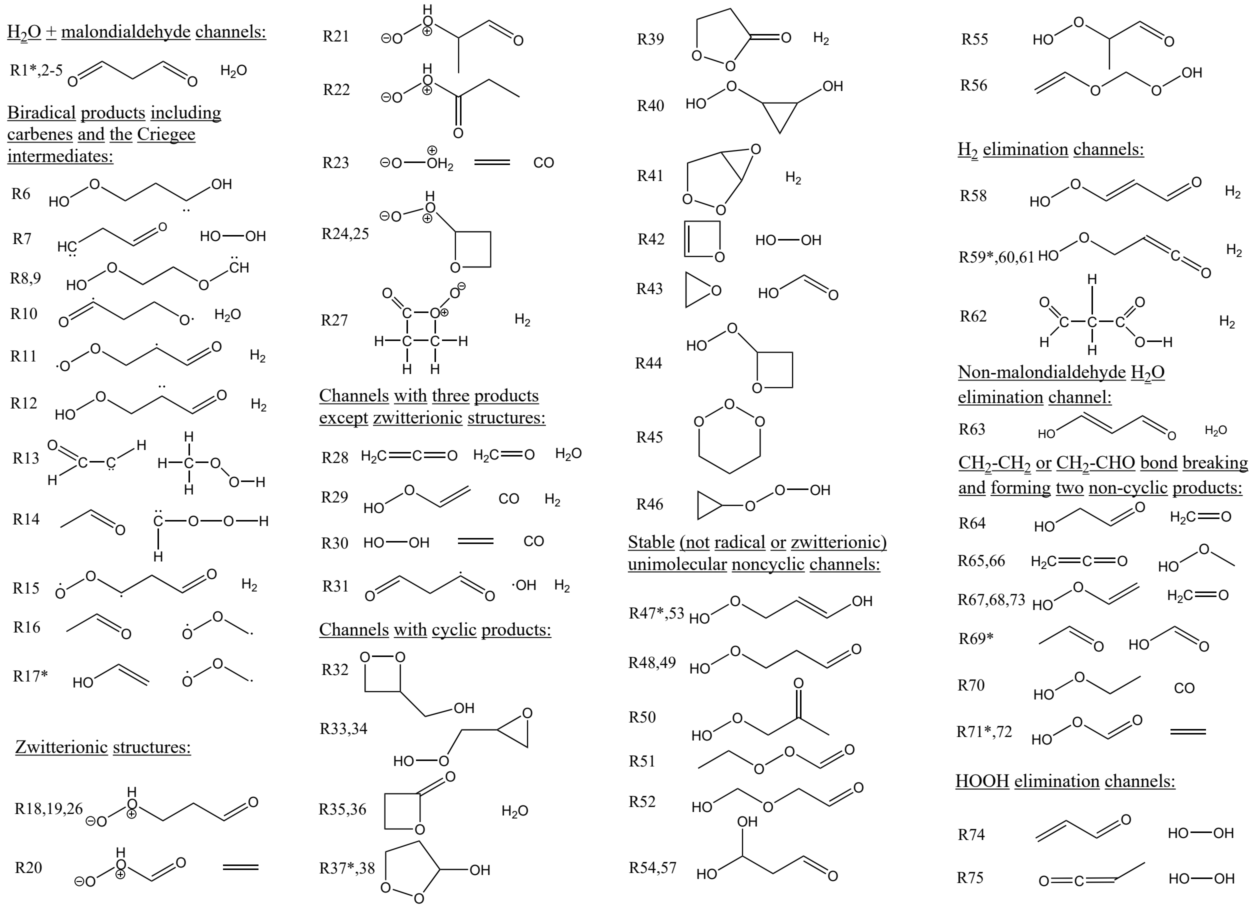

H2O + malondialdehyde channels:
R1*,2-5
Biradical products including carbenes and the Criegee intermediates:
R6
R7
R8,9
R10
R11
R12
R13
R14
R15
R16
R17*
Zwitterionic structures:
R18,19,26
R20
R21
R22
R23
R24,25
R27
Channels with three products except zwitterionic structures:
R28
R29
R30
R31
Channels with cyclic products:
R32
R33,34
R35,36
R37*,38
R39
R40
R41
R42
R43
R44
R45
R46
Stable (not radical or zwitterionic) unimolecular noncyclic channels:
R47*,53
R48,49
R50
R51
R52
R54,57
R55
R56
H2 elimination channels:
R58
R59*,60,61
R62
Non-malondialdehyde H2O elimination channel:
R63
CH2-CH2 or CH2-CHO bond breaking and forming two non-cyclic products:
R64
R65,66
R67,68,73
R69*
R70
R71*,72
HOOH elimination channels:
R74
R75

Figure 2.4. Automatically identified products of unimolecular reactions of γ-ketohydroperoxide (OOCCC=O) using the FSM, GSM, SSM, KinBot (first and second runs), and SC-AFIR methods. *Reactions discovered previously (reactions 1, 37, 47, 59, 69, 71 in Ref. [17] and reaction 17 in Ref. [53]).

2.4.1 Comparison with the existing chemical kinetic database

The present results also significantly outperform the existing RMG kinetic database where only four of the detected reactions (reactions 37, 40, 47, and 56) are present (Figure 2.5). In addition, RMG predicts 13 other chemical reactions which were not detected in the present study by any of the methods. Comparison between the present results and the RMG kinetic database suggests that the simple rules implemented in the RMG package may predict the correct bond breaking but they do not take into account additional internal H-transfers which stabilize the products. For instance, reaction 74 in Figure 2.4 has a similar RMG counterpart (RMG11 in Figure 2.5) but involves an additional H-transfer between the products, which stabilizes and reduces the overall product energy. In addition, the RMG library does not take into account channels with unusually long bond distances in the transition state that are detected in the present study (reactions 3, 4, 34, 36, 62) (note, however, that the bond lengths are not long enough for these reactions to be called roaming ones). Thus, the present work also suggests improvements to the RMG rules that

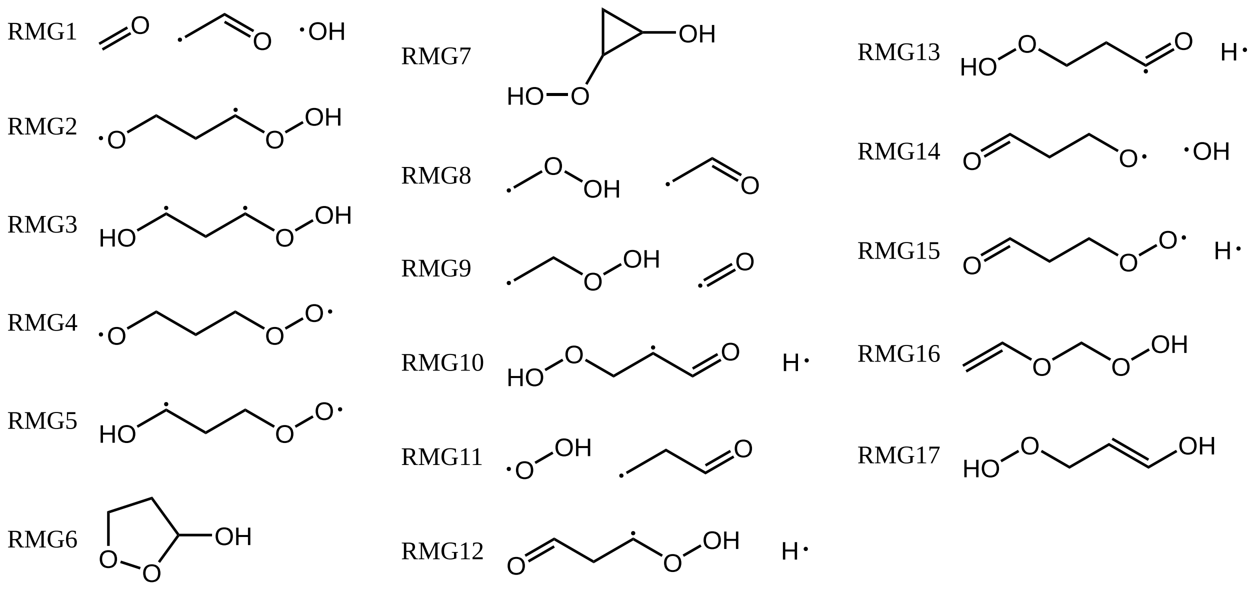


Figure 2.5. Unimolecular reactions of γ-ketohydroperoxide generated using the RMG kinetic database. RMG6 corresponds to R37, RMG7 to R40, RMG16 to R56, and RMG17 to R47.

could be implemented in the future. However, it is important to note that all of the methods in this study are primarily looking for saddle points, whereas some important dissociation channels (e.g., H elimination from the carbonyl group, see RMG13 in Figure 2.5) are barrierless (at least at the DFT level), therefore, the real number of reaction channels is always larger than the number of reactive saddle points for closed shell systems. For example, the Jalan et al. paper features the O−OH homolytic scission as the lowest barrierless channel, with a 49.5 kcal mol^{-1} barrier at the CCSD(T)//M06-2X level.[10] Our B3LYP calculations predict a $\sim$58–88 kcal mol^{-1} range for the water elimination channels. Given the tightness of these channels, the homolytic scission will dominate over the found saddle points. Moreover, some of the other reactions predicted by RMG involve additional biradical products, which are likely formed by high barrier processes that would have been excluded in the initial thermodynamic filtering for FSM/GSM/SSM and were also not found from suboptimal saddle point guesses that converged to unintended saddle points.

2.4.2 Types of identified chemical reactions

The reactions presented in Figure 2.4 are divided in the following way: H_2O + malondialdehyde channels (reactions 1–5); formation of biradical products including carbenes and Criegee intermediates (reactions 6–17); products with zwitterionic structures including (reactions 18–27); channels with three products except zwitterionic structures (reactions 28–31); channels with cyclic products (reactions 32–46); stable (not radical or zwitterionic) unimolecular noncyclic channels (reactions 47–57); H_2 elimination channels (reactions 58–62); another, non-malondialdehyde H_2O elimination channel (reaction 63); CH_2-CH_2 bond breaking and forming two noncyclic products (reactions

64–68); CHO – CH_2 bond breaking channels (reactions 69–73); HOOH elimination channels (reactions 74 and 75). Note that the reactions are grouped into these somewhat loosely defined groups primarily for orientation (for instance, the Criegee intermediates in reactions 15–17 can also be represented as their zwitterionic resonant structures). Also note that some of the detected reactions lead to the same product(s). For instance, we detected five different routes for H_2O elimination forming malondialdehyde (reactions 1–5). We labeled these as separate reaction channels, because their mechanisms are distinct, which would, for instance, gain importance in an experimental study using isotopically labeled KHP. We emphasize that our choice of considering such saddle points as fundamentally different is somewhat subjective especially since no rigorous conformational search was done to find the lowest energy conformers of the saddle points for any of the methods tested in the present work; however, channels marked as separate do not appear to be simple conformers of any other channels. It should be noted that SSM, SC-AFIR, and KinBot are capable of initiating multiple searches which could end in the same product because the product structure is not explicitly prescribed initially, whereas FSM and GSM only allow for one search per product. This enables SSM, SC-AFIR, and KinBot to find multiple distinct transition states corresponding to the same product. However, the initial set of reactions for SSM is the same as that for FSM and GSM in this study so that the methods can be compared directly in terms of their ability to find the same reactions. It is possible that this explains why KinBot was able to find all five routes leading to malondialdehyde, whereas none of the other methods were able to find all of those channels.

The majority of the detected reactions proceed through breaking two bonds (52), while the number of reactions with three bonds being broken is more than a factor of two smaller (21). Interestingly, we also observed one reaction with only one bond breakage (reaction 19 detected by all the methods) and one with four bond breakages (reaction 63 detected by FSM, SSM, and KinBot) even though none of these methods were targeted at reactions with four bond breakages. It is also worth noting that some of the detected product structures are highly unstable (e.g., the biradical structure in reaction 10 and the zwitterionic trivalent O structure in reaction 20) and therefore are expected to undergo fast subsequent reactions. Note that for these cases the nature of the product is sometimes very sensitive to the conformer of the saddle point and how the IRC is carried out (step lengths, etc.). Some conformations stabilize the biradicals, while others lead to ring-closures or further decomposition. For instance, the products of reaction 10 could rapidly undergo beta scission to yield the products corresponding to reaction 28. These reactions, if important kinetically, require further dynamical investigations (such as the analysis of plateau regimes of the real-time correlation function responsible for describing elementary chemical reactions in order to rigorously separate chemical events and detect the corresponding product channels), which lies outside the scope of this work. Some of the high-energy structures found by KinBot were a result of the reaction type extension in the second KinBot run. In the extended search, KinBot attempted to transfer all atom types to all other ones, which led to the additional discovery of reactions 8 and 18.

2.4.3 Exclusiveness of identified chemical reactions

Only four reactions were detected by all five methods: 1, 19, 37, and 74. If we exclude SC-AFIR, 18 reactions were detected by the remaining four methods (including FSM, GSM, SSM, and the first run of KinBot): 1, 2, 14, 15, 16, 19, 28, 30, 37, 47, 48, 51, 58, 64, 69, 71, 73, and 74. Including the results of the second KinBot run in the comparison adds three more reactions: 8, 13, and 39. Still, the reproducibility of the four methods is low in terms of the total number of channels found for each method. Except SC-AFIR, each method was able to detect some reactions exclusive to that method, although the majority of such saddle points are high in energy. The string methods detect several such exclusive reactions (7, 20, 35, and 72 from FSM; 22, 42, 43, and 65 from GSM; 23, 24, 38, 44, 68, and 75 from SSM). During the first run, KinBot also detects a few of them (5, 12, 17, 45, 66), but during the second "expanded" iteration the total number of exclusive reactions has more than doubled (13, 25, 27, 34, 46, 61, 62 in addition to the original set).

All five methods were able to detect the channel with the lowest energy barrier (approximately 35 kcal mol^{-1})—formation of 1,2-dioxolan-3-ol (reaction 37) which, as indicated in Section 2.1, is the first step in the so-called Korcek mechanism.[10] However, among the string methods, only GSM finds the Korcek reaction from the combinatorially intended product channel while FSM and SSM detect the corresponding saddle point from unintended channels. Interestingly, SSM was able to detect another enantiomer of 1,2-dioxolan-3-ol formed via the mechanism similar to the Korcek one (reaction 38), although its barrier is significantly higher (67.2 kcal mol^{-1}). KinBot detects the Korcek reaction as an internal 1,2-addition channel, a generic reaction from the first run. All methods except SC-AFIR were also able to detect the second lowest energy saddle point, for the reaction forming acetaldehyde and formic acid (reaction 69).

FSM often generates paths with multiple barriers (most noticeable reactions are 33, 37, 49, 62, 64, 70, and 72). In the previous study,[17] such profiles were discarded but the present results show that such cases can still lead to a discovery of the saddle points for single elementary steps (although typically with high energy barriers). Interestingly, in the previous study,[17] the Korcek reaction (reaction 37) was detected by FSM from the intended channel while the present FSM calculations detected it from the Berny optimization of the maximum along the string path for another channel, again confirming the seemingly random nature of the method and high sensitivity to the choice of input parameters.

SSM finds more intended reactions, however many of them represent a transfer of equivalent hydrogen atoms; only 9 out of 27 SSM channels are unique (see Table 2.1). This does not outnumber the FSM and GSM statistics by a lot, 7 out of 19 and 11 out of 24, respectively. Only three reactions were detected by all three string methods from the intended channels (reactions 9, 47, and 51) while the majority of saddle points detected by the string methods originate from unintended channels. Whether to consider this as a drawback of the proposed approach is not obvious. Many of the detected channels were unexpected and were not even included in the initial set of product channels (e.g., zwitterionic structures such as in reaction 19 where only one bond breaks and excessive valences are encountered, or reactions 48 and 49 where the products are isomorphic with the reactant molecule). However, our approach can easily fail if one has a specific reactant and product in mind. Because many saddle points were detected from unintended channels with string paths containing multiple elementary steps, it is expected that refining the initial set of channels to focus on more easily accessible single elementary steps would lead to a higher percentage of intended reactions, although this would limit the breadth of discovered reactions.

2.4.4 Chemical reactions with low-energy barriers and potentially important chemistry

Twenty-five chemical reactions with energy barriers below 40 $\mathrm{kcal\,mol^{-1}}$ (either forward or reverse channels) are summarized in Figure 2.6. Only two reactions (37 and 69) satisfy this criterion in the forward direction and both of them were detected in the previous FSM study,[17] suggesting that there is a low probability of finding additional low-energy channels (although not guaranteed). However, one can notice a significant number of previously unexpected reverse channels with low-energy barriers, many of which involve biradical and zwitterionic structures.

Several of the low-barrier reverse reactions leading to the formation of KHP involve Criegee intermediates as the reactant (reactions 15–17). These radicals play an important role in atmospheric chemistry and are in the focus of current intense research.[53–62] Reaction 17 involves the reaction of the simplest Criegee intermediate (CI) with vinyl alcohol (VA), which leads to KHP in a practically barrierless reaction. Vinyl alcohol is a detectable tautomer of acetaldehyde in flames[63] and in the atmosphere.[64,65] Interestingly, the reaction of the same CI with acetaldehyde (reaction 16) has a fairly noticeable barrier of 23 $\mathrm{kcal\,mol^{-1}}$ when the product is KHP. According to Jalan et al., CI can also react with acetaldehyde in a barrierless manner, but such a reaction leads to a secondary ozonide.[62] Another pair of low-barrier chemical reactions presented in Figure 2.6 is H_2 fission by a 3-carbon Criegee intermediate (reaction 15) with $E_a = 16\,\mathrm{kcal\,mol^{-1}}$ and a reaction involving a similar biradical structure (reaction 11) which exhibits a much smaller barrier of 5 $\mathrm{kcal\,mol^{-1}}$. However, reactions of CI with H_2 are likely not of relevance in the atmosphere due to the low concentration of hydrogen molecules.[66]

The zwitterionic structures detected in reactions 18–27 contain hypervalent O-atoms, which made their discovery highly surprising. They were not included in the product set of the string methods used in the present study due to their unusual valency and were thus discovered serendipitously. However, because KinBot does not consider products to guide its search, the H-transfer to the inner O atom of the $-$OOH groups was an allowed step. It is notable that six out of ten detected reactions of these zwitterionic structures fell within our category of "low-energy" barriers. Reac-

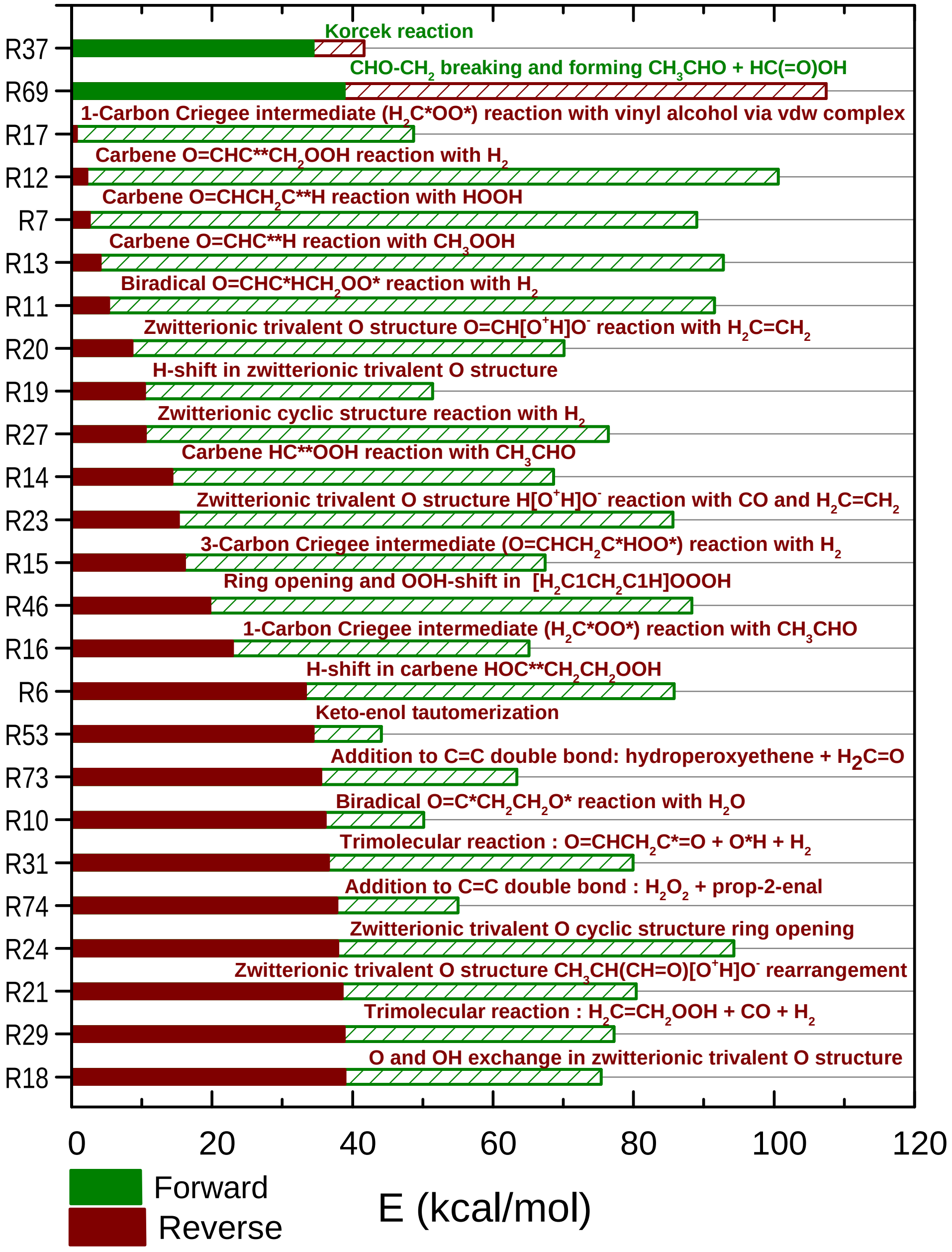


Figure 2.6. Energy barriers (E_a) of 25 chemical reactions of unimolecular decomposition and isomerization of γ-ketohydroperoxide for which E_a is below $40\,\mathrm{kcal\,mol^{-1}}$ (either for the forward or reverse channel). Note that bars representing the barrier heights are not stacked but overlapped for each reaction.

tion 23 forms water oxide[67,68] and the other channels form ylides of varying complexity. Although usually unstable in the gas phase, certain subgroups of zwitterions often appear in biochemical studies (under physiological conditions, a variety of biomolecules, such as amino acids, exist as zwitterions) and in organic synthesis. The present results particularly contribute to the chemical knowledge for zwitterionic structures with positively and negatively charged atoms in the O_2

fragment. This has potential implications on the chemistry of alkylperoxides, a class of species important in low-temperature autoignition and many atmospheric chemistry problems. For instance, it has been shown that water elimination from alkylperoxy radicals can involve zwitterionic states.[11] Apart from water oxide, the ylides discovered by the automated search have not been analyzed in much detail in literature. Schalley et al. conducted an extensive characterization of the PESs of methanol oxide and dimethyl ether oxide[69] and Vereecken et al. recently discovered the relevance of alcohol and ether oxides in atmospheric chemistry systems involving a Criegee intermediate.[70] Alcohol oxides, such as those in reactions 18, 19, 21, 24, 25, and 26 are expected to undergo rapid tautomerization to the corresponding hydroperoxides, whereas ether oxides are more stable.[69,70] In particular, reaction 27 involves a cyclic ether oxide for which isomerization to an alkylperoxide is expected to be associated with a very significant energy barrier indicating that the species could be long-lived enough to participate in other reactions, although it is not clear in which chemical systems such a species would be present in significant amounts. To the best of our knowledge, carboxylic acid oxides, such as those in reactions 20 and 22, have not been reported in literature, and the low barrier of the reaction between formic acid oxide and ethylene indicates that these species could potentially be relevant if the tautomerization to the corresponding peroxy acid is slow enough.

Another interesting class of structures, carbenes, was detected in reactions 12, 7, 13, 14, 6 (sorted according to the height of the corresponding energy barriers in Figure 2.6). These reactions were not expected as the forward reactions are associated with large energy barriers. The version of group additivity used for the creation of the filtered data set mostly underestimates the endothermicity for such reactions compared to more accurate B3LYP energies, which led to their inclusion in the set of filtered reactions. Reaction 6 corresponds to an intramolecular O−H insertion (or internal H-shift), while the remaining four reactions are intermolecular insertions. Both are common types of carbene reactions; however, the context in which they were discovered is surprising. The observed "kinetic" preference of carbene insertion into an existing bond is in accordance with the common rule for carbene X−H insertion: Reactions 12 (X = H) and 7 (X = O) have lower energy barriers than reactions 13 and 14 (X = C). The intramolecular carbene insertion reaction 8 (O−C) was not included in Figure 2.6 due to an energy barrier higher than 40 kcal mol^{-1}. Our results show that the carbene insertions occur in single elementary steps, which is possible because we are only considering the singlet surface for this study,[71] whereas triplet carbenes would have to react in a sequence of two steps. Reaction 7 is notable due to the reaction with hydrogen peroxide, which is present in most combustion models and therefore might constitute a valid path to KHP. Even if the carbene is only present for short amount of times, its singlet state will result in high rate coefficients, and might make such a reaction important. Cyclopropene, or its isomers allene and propyne, is produced in a reaction between methylene and acetylene,[72,73] and is relevant in combustion systems.[74] It is conceivable for oxygen atoms to add to the double bond in cyclopropene, which could then undergo further ring-opening to the carbene in reaction 7. Alternatively, it would be possible for oxygen to react directly with acetylene, the product of which could then react with methylene and also lead to the carbene. We note that such thoughts are purely hypothetical at this point, but they may warrant further investigation, which could show that species like the carbene in reaction 7 are important in combustion systems. Similarly, reaction 13 involves a carbene that might be formed from vinoxy radicals, which are also important in combustion systems,[75] and thus may be a relevant reaction in such systems.

Although the low-barrier chemical reactions described above are potentially important, calculating rate coefficients for all of the reactions is not the goal of the present work. However, we decided to investigate one reaction (17) in more detail as it involves promising new chemistry that may be relevant in atmospheric systems. A characterization of KHP formation resulting from the reaction of CI with VA, reaction 17, was published by Vereecken during the preparation of this manuscript.[53] In accordance with our findings, Vereecken found that the reaction proceeds through a van der Waals complex and a subsequent shallow barrier. Apart from the brief mention in Ref. [53], we are not aware of any other publications that have studied this reaction in greater depth, even though its small submerged barrier suggests a fast rate, which may render it important in atmospheric systems.

In general, tautomerization of acetaldehyde to VA is associated with a high barrier in the gas phase, although it has been discovered that photo-tautomerization of acetaldehyde is a viable process in the atmosphere, which can lead to up to 21% stable VA produced from excited acetaldehyde.[65,76] Due to the large energy of CI and VA relative to KHP, it should be possible to skip the KHP well at finite pressures and to obtain other bimolecular decomposition products of KHP directly. Likely candidates are the products obtained from the dissociation of the oxygen single bond in the hydroperoxy group of KHP, which yields two radicals (note that such products were not included in the search space for the automated TS search methods because the methods are not suitable for purely barrierless processes). To obtain quantitatively accurate results, we reoptimized the relevant geometries and calculated frequencies at the UM06-2X/aug-cc-pVTZ level of theory with an ultrafine integration grid and tight SCF convergence criteria. We subsequently calculated energies with the RCCSD(T)-F12a/cc-pVTZ-F12 method. The PES for the CI + VA network is shown in Figure 2.7.

The Arkane code, distributed as part of RMG,[44] was used for all kinetic rate calculations. One-dimensional hindered rotor scans were completed for all rotatable bonds to enable accurate calculation of relevant partition functions. Frequencies were scaled by a factor of 0.971, consistent with Ref. [77]. For pressure-dependent calculations, the one-dimensional master equation was solved using the Reservoir State approximation employing an exponential down collisional energy-transfer model with nitrogen as the bath gas and using an average downward collisional energy transfer of $150\,\mathrm{cm}^{-1}$ and an active K-rotor and inactive J-rotor.[65] Lennard-Jones parameters were estimated using the RMG website (https://rmg.mit.edu/). The rate of formation of the van der Waals complex was not assumed to affect the overall rate, the justification for which is shown in a more in-depth analysis in Ref. [78]. The rate for the barrierless dissociation of KHP to the two radical products was estimated from an analogous reaction of ethyl hydroperoxide.[79]

The rate constant for the reaction of CI and VA was calculated as 3.7×10^{-11} cm^3 molecule^{-1} s^{-1} at 298 K, which is faster or on the same order of other dominant reactions of the Criegee intermediate.[80] At atmospheric conditions, virtually all of the flux is towards the two radical products including OH and none of the product is obtained in the KHP form, which indicates that this reaction may be a significant source of OH and alkoxy radicals in the atmosphere and could help explain the underprediction of OH in current models compared to experiment.[81,82] However, it should be noted that the majority of Criegee intermediates available for reaction with VA in the atmosphere have larger substituent groups as a result of easier collisional stabilization after their production from larger primary ozonides.[81] Therefore, thermalization to KHP will be more likely for larger CI. *A priori*, it is conceivable for CI to add to the double bond in VA instead of forming KHP, which can lead to two separate ring structures depending on the orientation during the addition reaction (in fact, one of the structures is identical to the product of the first step of the Korcek mechanism). Rate calculations for these reactions showed that they are not relevant at atmospheric conditions. Reaction with OH and O_2 constitutes the main reaction of VA in the atmosphere with a total rate constant of 6.8×10^{-11} cm^3 molecule^{-1} s^{-1} at 298 K.[83] The concentration of stabilized CI in the atmosphere is associated with significant uncertainty and ranges from approximately 10^3 to 5.5×10^4 molecule cm^{-3} for global average concentrations and has peak concentrations of 10^5 molecule cm^{-3}.[80,84] Global average OH concentration can be estimated as 10^6 molecule cm^{-3}.[83] This suggests that the branching ratio between the reaction of CI and VA and the reaction of OH and VA is approximately 0.05 at peak CI conditions, although this estimate is quite variable based on uncertainties in the concentrations and in the calculated rates.

As it is beyond the scope of the current study, we did not evaluate the effect of different substituents in CI, which are known to strongly affect rate constants.[80] Such an analysis may lead to more competitive rates of the CI + VA reaction. Enols other than vinyl alcohol may also play a role in atmospheric chemistry and could also undergo reaction with CI in a similar process. For example, acetylacetone exists predominantly in its enol form and is expected to be present in the atmosphere due to its widespread use in industry and ensuing release.[85] Such stable enols may be especially prone to reacting with Criegee intermediates, as photochemical activation is not required for their production, which may even render them important in nighttime conditions. Therefore, a subsequent study analyzing both the effect of substituents in CI and the reactivity towards stable enols would be of significance.

2.5 Conclusions

Transition states (TSs) for 68 previously unknown reactions of the simplest γ-ketohydroperoxide were discovered automatically by combining five different saddle-point-search algorithms. This sig-

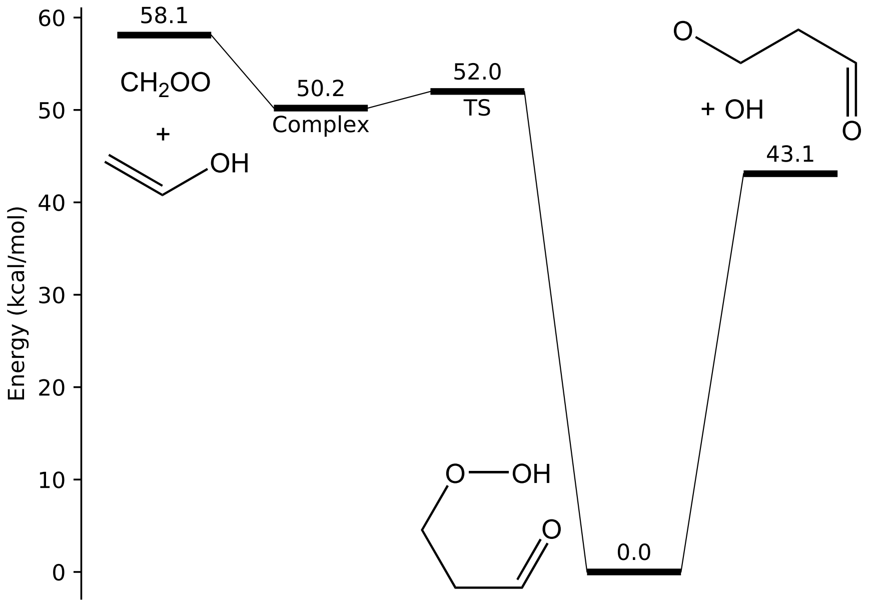


Figure 2.7. Potential energy surface of the Criegee + vinyl alcohol reaction at the RCCSD(T)-F12a/cc-pVTZ-F12//UM06-2X/aug-cc-pVTZ level of theory, which proceeds through a barrierless formation of a van der Waals complex followed by a shallow transition state to KHP, which can dissociate to two radicals. Well-skipping directly to the radical products is the preferred channel at atmospheric pressures. All values include the zero-point energy.

nificantly broadens the chemical knowledge for the present system and for all γ-ketohydroperoxides. The single-ended growing string method (SSM) found TSs leading to 50 distinct products, the most of all methods tested, while the single-component artificial force induced reaction method (SC-AFIR) detected the fewest (7). The remaining methods demonstrated intermediate results: TSs to 39 product channels were discovered by the freezing string method (FSM), 46 were discovered by the double-ended growing string method (GSM), and 32 were found by KinBot (generic run). The present FSM results outperform the previous ones,[17] where only six reaction channels were detected using the same method, showing that the overall performance of the method is highly dependent on the input parameters and the internal details of the computational procedure. All of the methods were able to detect the reaction with the lowest energy barrier, which corresponds to the first step of the so-called Korcek mechanism.[10] However, among the string methods, only GSM detected the Korcek reaction when searching for this channel, while FSM and SSM found this saddle point serendipitously while searching for TSs corresponding to different products. All methods, except for SC-AFIR, detected several exclusive channels (i.e., reactions detected only by a given method). SC-AFIR found the highest percentage of conformational saddle points.

Analysis of the reverse reactions leading to the formation of γ-ketohydroperoxide reveals promising new chemistry with low energy barriers for reactions involving zwitterions, biradicals, including carbenes, and Criegee intermediates. Similar species are in the focus of current intense research and we show that several of the detected reactions may be relevant in atmospheric and combustion chemistry. In particular, we study the reaction of the simplest Criegee intermediate with vinyl alcohol in more detail, which leads to γ-ketohydroperoxide via a submerged barrier reaction with possible well-skipping to two radical products, by calculating the rate constant for this reaction at atmospheric conditions and comparing it to other possible decay processes of vinyl alcohol. As a result of our analysis, directions for subsequent chemical rate studies have been formulated exhibiting a clear example of the benefits of applying automated TS search algorithms for the discovery of new chemical reactions.

The present results highlight the complexity of chemical space exploration and the sensitivity to input parameters and level of theory of the methods assessed in the present work. They also demonstrate the power of joint application of several automated approaches for discovery of elementary chemical reactions since none of the above listed methods is greatly superior to the others and all methods discover TSs not detected by any other algorithm (except SC-AFIR). Moreover, chemical knowledge obtained using one or more methods can be used to improve the performance of another method. For example, during the second iteration of the KinBot calculations, selected due to its flexibility and the ease of implementing modifications, the reaction types were extended using the chemical knowledge obtained in the first run by all five methods resulting in significant improvement; TSs leading to 16 additional product channels were detected by the extended version of KinBot. At the same time, such results also show that it is impossible to determine how exhaustive an automated search was, as none of the methods are capable of finding all of the reactions and there is no guarantee that all reactions can be discovered by the joint approach.

Our analysis also shows various sources of crashed searches and excessive (useless or repetitive) computation for the methods selected for the present comparative study. We hope this information can be used to improve future applications of the methods. In the future, we plan to extend the present work to similar algorithms based on other representative methods for saddle point detection, such as the nudged elastic band method,[28,29] conjugate peak refinement,[86] the ridge method,[87] and others. We envision that similar method assessment for bimolecular chemical reactions would be a natural subsequent step for the community.

The results of this study indicate a vast number of unexpected elementary-step chemical reactions remain to be discovered, and that despite the weaknesses of existing TS search algorithms, this reaction discovery process can be greatly accelerated by automated search of potential energy surfaces.

2.6 References

(1) Martínez, T. J. Ab Initio Reactive Computer Aided Molecular Design. *Acc. Chem. Res.* **2017**, *50*, 652–656.

(2) Dewyer, A. L.; Zimmerman, P. M. Finding Reaction Mechanisms, Intuitive or Otherwise. *Org. Biomol. Chem.* **2017**, *15*, 501–504.

(3) Klippenstein, S. J. From Theoretical Reaction Dynamics to Chemical Modeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77–111.

(4) Vinu, R.; Broadbelt, L. J. Unraveling Reaction Pathways and Specifying Reaction Kinetics for Complex Systems. *Annu. Rev. Chem. Biomol. Eng.* **2012**, *3*, 29–54.

(5) Schlögl, R. Heterogeneous Catalysis. *Angew. Chem. Int. Ed.* **2015**, *54*, 3465–3520.

(6) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017**, *50*, 605–608.

(7) Hammes-Schiffer, S. Catalysts by Design: The Power of Theory. *Acc. Chem. Res.* **2017**, *50*, 561–566.

(8) Vereecken, L.; Glowacki, D. R.; Pilling, M. J. Theoretical Chemical Kinetics in Tropospheric Chemistry: Methodologies and Applications. *Chem. Rev.* **2015**, *115*, 4063–4114.

(9) Van de Vijver, R.; Devocht, B. R.; Van Geem, K. M.; Thybaut, J. W.; Marin, G. B. Challenges and Opportunities for Molecule-Based Management of Chemical Processes. *Curr. Opin. Chem. Eng.* **2016**, *13*, 142–149.

(10) Jalan, A.; Alecu, I. M.; Meana-Pañeda, R.; Aguilera-Iparraguirre, J.; Yang, K. R.; Merchant, S. S.; Truhlar, D. G.; Green, W. H. New Pathways for Formation of Acids and Carbonyl Products in Low-Temperature Oxidation: The Korcek Decomposition of γ-Ketohydroperoxides. *J. Am. Chem. Soc.* **2013**, *135*, 11100–11114.

(11) Welz, O.; Klippenstein, S. J.; Harding, L. B.; Taatjes, C. A.; Zádor, J. Unconventional Peroxy Chemistry in Alcohol Oxidation: The Water Elimination Pathway. *J. Phys. Chem. Lett.* **2013**, *4*, 350–354.

(12) Quelch, G. E.; Gallo, M. M.; Schaefer, H. F. Aspects of the Reaction Mechanism of Ethane Combustion. Conformations of the Ethylperoxy Radical. *J. Am. Chem. Soc.* **1992**, *114*, 8239–8247.

(13) Ludwig, J. R.; Zimmerman, P. M.; Gianino, J. B.; Schindler, C. S. Iron(III)-Catalysed Carbonyl-Olefin Metathesis. *Nature* **2016**, *533*, 374–379.

(14) Khomutnyk, Y. Y.; Argüelles, A. J.; Winschel, G. A.; Sun, Z.; Zimmerman, P. M.; Nagorny, P. Studies of the Mechanism and Origins of Enantioselectivity for the Chiral Phosphoric Acid-Catalyzed Stereoselective Spiroketalization Reactions. *J. Am. Chem. Soc.* **2016**, *138*, 444–456.

(15) Magoon, G. R.; Green, W. H. Design and Implementation of a Next-Generation Software Interface for On-The-Fly Quantum and Force Field Calculations in Automated Reaction Mechanism Generation. *Comput. Chem. Eng.* **2013**, *52*, 35–45.

(16) Bera, P. P.; Sattelmeyer, K. W.; Saunders, M.; Schaefer III, H. F.; Schleyer, P. R. Mindless Chemistry. *J. Phys. Chem. A* **2006**, *110*, 4287–4290.

(17) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(18) Yu, T.; Zheng, J.; Truhlar, D. G. Multipath Variational Transition State Theory: Rate Constant of the 1,4-Hydrogen Shift Isomerization of the 2-Cyclohexylethyl Radical. *J. Phys. Chem. A* **2012**, *116*, 297–308.

(19) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(20) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry with an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6*, 1044–1048.

(21) Maeda, S.; Harabuchi, Y.; Takagi, M.; Taketsugu, T.; Morokuma, K. Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces. *Chem. Rec.* **2016**, *16*, 2232–2248.

(22) Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

(23) Yang, M.; Zou, J.; Wang, G.; Li, S. Automatic Reaction Pathway Search via Combined Molecular Dynamics and Coordinate Driving Method. *J. Phys. Chem. A* **2017**, *121*, 1351–1361.

(24) Zimmerman, P. Reliable Transition State Searches Integrated with the Growing String Method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(25) Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722.

(26) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient Exploration of Reaction Paths via a Freezing String Method. *J. Chem. Phys.* **2011**, *135*, 224108.

(27) Zimmerman, P. M. Growing String Method with Interpolation and Optimization in Internal Coordinates: Method and Examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(28) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(29) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(30) Bhoorasingh, P.; West, R. Transition State Geometry Prediction Using Molecular Group Contributions. *Phys. Chem. Chem. Phys.* **2015**, *17*, 32173–32182.

(31) Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **1982**, *3*, 214–218.

(32) Peng, C. Y.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States. *J. Comput. Chem.* **1996**, *17*, 49–56.

(33) Schlegel, H. B. Estimating the Hessian for Gradient Type Geometry Optimizations. *Theor. Chim. Acta* **1984**, *66*, 333–340.

(34) Zádor, J.; Najm, H. N., *KinBot 1.0: A Code for Automatic PES Exploration*; Sandia National Laboratories: Livermore, CA, 2013.

(35) Van de Vijver, R.; Zádor, J. KinBot: Automated Stationary Point Search on Potential Energy Surfaces. *Comput. Phys. Commun.* **2020**, *248*, 106947.

(36) Zheng, J.; Mielke, S. L.; Clarkson, K. L.; Truhlar, D. G. MSTor: A Program for Calculating Partition Functions, Free Energies, Enthalpies, Entropies, and Heat Capacities of Complex Molecules Including Torsional Anharmonicity. *Comput. Phys. Commun.* **2012**, *183*, 1803–1812.

(37) Zheng, J.; Meana-Pañeda, R.; Truhlar, D. G. MSTor Version 2013: A New Version of the Computer code for the Multi-Structural Torsional Anharmonicity, Now with a Coupled Torsional Potential. *Comput. Phys. Commun.* **2013**, *184*, 2032–2033.

(38) Yu, T.; Zheng, J.; Truhlar, D. G. Multi-Structural Variational Transition State Theory. Kinetics of the 1,4-Hydrogen Shift Isomerization of the Pentyl Radical with Torsional Anharmonicity. *Chem. Sci.* **2011**, *2*, 2199–2213.

(39) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.

(40) Sameera, W. M.; Maeda, S.; Morokuma, K. Computational Catalysis Using the Artificial Force Induced Reaction Method. *Acc. Chem. Res.* **2016**, *49*, 763–773.

(41) Jensen, R. K.; Korcek, S.; Mahoney, L. R.; Zinbo, M. Liquid-Phase Autoxidation of Organic Compounds at Elevated Temperatures. 1. The Stirred Flow Reactor Technique and Analysis of Primary Products from n-Hexadecane Autoxidation at 120–180 °C. *J. Am. Chem. Soc.* **1979**, *101*, 7574–7584.

(42) Jensen, R. K.; Korcek, S.; Mahoney, L. R.; Zinbo, M. Liquid-Phase Autoxidation of Organic Compounds at Elevated Temperatures. 2. Kinetics and Mechanisms of the Formation of Cleavage Products in n-Hexadecane Autoxidation. *J. Am. Chem. Soc.* **1981**, *103*, 1742–1749.

(43) Zádor, J.; Taatjes, C. A.; Fernandes, R. X. Kinetics of Elementary Reactions in Low-Temperature Autoignition Chemistry. *Prog. Energy Combust. Sci.* **2011**, *37*, 371–421.

(44) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(45) Benson, S. W., *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*, 2nd edition; Wiley: New York, 1976.

(46) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*.

(47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision C.01, Wallingford, CT, 2016.

(48) Zimmerman, P. molecularGSM. https://github.com/ZimmermanGroup/molecularGSM (accessed 08/10/2017).

(49) Maeda, S.; Taketsugu, T.; Morokuma, K. Exploring Transition State Structures for Intramolecular Pathways by the Artificial Force Induced Reaction Method. *J. Comput. Chem.* **2014**, *35*, 166–173.

(50) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(51) Walker, M.; Harvey, A. J.; Sen, A.; Dessent, C. E. Performance of M06, M06-2X, and M06-HF Density Functionals for Conformationally Flexible Anionic Clusters: M06 Functionals Perform Better than B3LYP for a Model System with Dispersion and Ionic Hydrogen-Bonding Interactions. *J. Phys. Chem. A* **2013**, *117*, 12590–12600.

(52) Maeda, S.; Harabuchi, Y. On Benchmarking of Automated Methods for Performing Exhaustive Reaction Path Search. *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

(53) Vereecken, L. The Reaction of Criegee Intermediates with Acids and Enols. *Phys. Chem. Chem. Phys.* **2017**, *19*, 28630–28640.

(54) Welz, O.; Savee, J. D.; Osborn, D. L.; Vasu, S. S.; Percival, C. J.; Shallcross, D. E.; Taatjes, C. A. Direct Kinetic Measurements of Criegee Intermediate (CH_2OO) Formed by Reaction of CH_2I with O_2. *Science* **2012**, *335*, 204–207.

(55) Taatjes, C. A.; Welz, O.; Eskola, A. J.; Savee, J. D.; Scheer, A. M.; Shallcross, D. E.; Rotavera, B.; Lee, E. P. F.; Dyke, J. M.; Mok, D. K. W.; Osborn, D. L.; Percival, C. J. Direct Measurements of Conformer-Dependent Reactivity of the Criegee Intermediate CH_3CHOO. *Science* **2013**, *340*, 177–180.

(56) Taatjes, C. A.; Welz, O.; Eskola, A. J.; Savee, J. D.; Osborn, D. L.; Lee, E. P.; Dyke, J. M.; Mok, D. W.; Shallcross, D. E.; Percival, C. J. Direct Measurement of Criegee Intermediate (CH_2OO) Reactions with Acetone, Acetaldehyde, and Hexafluoroacetone. *Phys. Chem. Chem. Phys.* **2012**, *14*, 10391–10400.

(57) Tobias, H. J.; Ziemann, P. J. Kinetics of the Gas-Phase Reactions of Alcohols, Aldehydes, Carboxylic Acids, and Water with the C13 Stabilized Criegee Intermediate Formed from Ozonolysis of 1-Tetradecene. *J. Phys. Chem. A* **2001**, *105*, 6129–6135.

(58) Taatjes, C. A.; Meloni, G.; Selby, T. M.; Trevitt, A. J.; Osborn, D. L.; Percival, C. J.; Shallcross, D. E. Direct Observation of the Gas-Phase Criegee Intermediate (CH_2OO). *J. Am. Chem. Soc.* **2008**, *130*, 11883–11885.

(59) Andersen, A.; Carter, E. A. Hybrid Density Functional Theory Predictions of Low-Temperature Dimethyl Ether Combustion Pathways. II. Chain-Branching Energetics and Possible Role of the Criegee Intermediate. *J. Phys. Chem. A* **2003**, *107*, 9463–9478.

(60) Buras, Z. J.; Elsamra, R. M.; Green, W. H. Direct Determination of the Simplest Criegee Intermediate (CH 2OO) Self Reaction Rate. *J. Phys. Chem. Lett.* **2014**, *5*, 2224–2228.

(61) Buras, Z. J.; Elsamra, R. M.; Jalan, A.; Middaugh, J. E.; Green, W. H. Direct Kinetic Measurements of Reactions Between the Simplest Criegee Intermediate CH_2OO and Alkenes. *J. Phys. Chem. A* **2014**, *118*, 1997–2006.

(62) Jalan, A.; Allen, J. W.; Green, W. H. Chemically Activated Formation of Organic Acids in Reactions of the Criegee Intermediate with Aldehydes and Ketones. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16841.

(63) Taatjes, C. A.; Hansen, N.; McIlroy, A.; Miller, J. A.; Senosiain, J. P.; Klippenstein, S. J.; Qi, F.; Sheng, L.; Zhang, Y.; Cool, T. A.; Wang, J.; Westmoreland, P. R.; Law, M. E.; Kasper, T.; Kohse-Höinghaus, K. Enols are Common Intermediates in Hydrocarbon Oxidation. *Science* **2005**, *308*, 1887–1889.

(64) Heazlewood, B. R.; MacCarone, A. T.; Andrews, D. U.; Osborn, D. L.; Harding, L. B.; Klippenstein, S. J.; Jordan, M. J. T.; Kable, S. H. Near-Threshold H/D Exchange in CD_3CHO Photodissociation. *Nat. Chem.* **2011**, *3*, 443–448.

(65) Andrews, D. U.; Heazlewood, B. R.; Maccarone, A. T.; Conroy, T.; Payne, R. J.; Jordan, M. J. T.; Kable, S. H. Photo-Tautomerization of Acetaldehyde. *Science* **2012**, *337*, 1203–1206.

(66) Novelli, P. C.; Lang, P. M.; Masarie, K. A.; Hurst, D. F.; Myers, R.; Elkins, J. W. Molecular Hydrogen in the Troposphere: Global Distribution and Budget. *J. Geophys. Res.* **1999**, *104*, 30427–30444.

(67) Schröder, D.; Schalley, C. A.; Goldberg, N.; Hrusak, J.; Schwarz, H. Gas-Phase Experiments Aimed at Probing the Existence of the Elusive Water Oxide Molecule. *Angew. Chem. Int. Ed.* **1996**, *35*, 1235–1242.

(68) Bach, R. D.; Owensby, A. L.; Gonzalez, C.; Bernhard Schlegel, H.; McDouall, J. J. Nature of the Transition Structure for Oxygen Atom Transfer from a Hydroperoxide. Theoretical Comparison between Water Oxide and Ammonia Oxide. *J. Am. Chem. Soc.* **1991**, *113*, 6001–6011.

(69) Schalley, C. A.; Harvey, J. N.; Schröder, D.; Schwarz, H. Ether Oxides: A New Class of Stable Ylides? A Theoretical Study of Methanol Oxide and Dimethyl Ether Oxide. *J. Phys. Chem. A* **1998**, *102*, 1021–1035.

(70) Vereecken, L.; Rickard, A. R.; Newland, M. J.; Bloss, W. J. Theoretical Study of the Reactions of Criegee Intermediates with Ozone, Alkylhydroperoxides, and Carbon Monoxide. *Phys. Chem. Chem. Phys.* **2015**, *17*, 23847–23858.

(71) Bach, R. D.; Su, M. D.; Aldabbagh, E.; Andrés, J. L.; Schlegel, H. B. A Theoretical Model for the Orientation of Carbene Insertion into Saturated Hydrocarbons and the Origin of the Activation Barrier. *J. Am. Chem. Soc.* **1993**, *115*, 10237–10246.

(72) Yu, H. G.; Muckerman, J. T. Ab Initio and Direct Dynamics Studies of the Reaction of Singlet Methylene with Acetylene and the Lifetime of the Cyclopropene Complex. *J. Phys. Chem. A* **2005**, *109*, 1890–1896.

(73) Frankcombe, T. J.; Smith, S. C. Time-Dependent Master Equation Simulation of Complex Elementary Reactions in Combustion: Application to the Reaction of $^1CH_2$ with C_2H_2 from 300–2000 K. *Faraday Discuss.* **2001**, *119*, 159–171.

(74) Blanquart, G.; Pepiot-Desjardins, P.; Pitsch, H. Chemical Mechanism for High Temperature Combustion of Engine Relevant Fuels with Emphasis on Soot Precursors. *Combust. Flame* **2009**, *156*, 588–607.

(75) Senosiain, J. P.; Klippenstein, S. J.; Miller, J. A. Pathways and Rate Coefficients for the Decomposition of Vinoxy and Acetyl Radicals. *J. Phys. Chem. A* **2006**, *110*, 5772–5781.

(76) Clubb, A. E.; Jordan, M. J. T.; Kable, S. H.; Osborn, D. L. Phototautomerization of Acetaldehyde to Vinyl Alcohol: A Primary Process in UV-Irradiated Acetaldehyde from 295 to 335 nm. *J. Phys. Chem. Lett.* **2012**, *3*, 3522–3526.

(77) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(78) Elsamra, R. M.; Jalan, A.; Buras, Z. J.; Middaugh, J. E.; Green, W. H. Temperature- and Pressure-Dependent Kinetics of CH_2OO + CH_3COCH_3 and CH_2OO + CH_3CHO: Direct Measurements and Theoretical Analysis. *Int. J. Chem. Kinet.* **2016**, *48*, 474–488.

(79) Chen, D.; Jin, H.; Wang, Z.; Zhang, L.; Qi, F. Unimolecular Decomposition of Ethyl Hydroperoxide: Ab Initio/Rice-Ramsperger-Kassel-Marcus Theoretical Prediction of Rate Constants. *J. Phys. Chem. A* **2011**, *115*, 602–611.

(80) Vereecken, L.; Novelli, A.; Taraborrelli, D. Unimolecular Decay Strongly Limits the Atmospheric Impact of Criegee Intermediates. *Phys. Chem. Chem. Phys.* **2017**, *19*, 31599–31612.

(81) Vereecken, L.; Francisco, J. S. Theoretical Studies of Atmospheric Reaction Mechanisms in the Troposphere. *Chem. Soc. Rev.* **2012**, *41*, 6259–6293.

(82) Stone, D.; Whalley, L. K.; Heard, D. E. Tropospheric OH and HO_2 Radicals: Field Measurements and Model Comparisons. *Chem. Soc. Rev.* **2012**, *41*, 6348.

(83) So, S.; Wille, U.; Da Silva, G. Atmospheric Chemistry of Enols: A Theoretical Study of the Vinyl Alcohol + OH + O_2 Reaction Mechanism. *Environ. Sci. Technol.* **2014**, *48*, 6694–6701.

(84) Novelli, A.; Hens, K.; Ernest, C. T.; Martinez, M.; Nölscher, A. C.; Sinha, V.; Paasonen, P.; Petäjä, T.; Sipilä, M.; Elste, T.; Plass-Dülmer, C.; Phillips, G. J.; Kubistin, D.; Williams, J.; Vereecken, L.; Lelieveld, J.; Harder, H. Estimating the Atmospheric Concentration of Criegee Intermediates and Their Possible Interference in a FAGE-LIF Instrument. *Atmos. Chem. Phys.* **2017**, *17*, 7807–7826.

(85) Zhou, S.; Barnes, I.; Zhu, T.; Bejan, J.; Albu, M.; Benter, T. Atmospheric Chemistry of Acetylacetone. *Environ. Sci. Technol.* **2008**, *42*, 7905–7910.

(86) Fischer, S.; Karplus, M. Conjugate Peak Refinement: an Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chem. Phys. Lett.* **1992**, *194*, 252–261.

(87) Ionova, I. V.; Carter, E. A. Ridge Method for Finding Saddle Points on Potential Energy Surfaces. *J. Chem. Phys.* **1993**, *98*, 6377–6386.

Chapter 3

Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach

Much of this work has previously appeared as Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835. Yi-Pei Li wrote significant parts of the code. Allen Mark Payne was involved in developing the automated framework for fitting bond additivity corrections described in Section 3.6. The code for training the machine learning models is available at `https://github.com/ReactionMechanismGenerator/DataDrivenEstimator` and the bond additivity correction code is distributed in the Arkane package as part of RMG available at `https://github.com/ReactionMechanismGenerator/RMG-Py`.

3.1 Introduction

Rapid and accurate estimation of molecular properties is a vital component of many chemistry and materials science applications.[1,2] In particular, automated reaction mechanism generation requires estimates of chemical kinetic rates and molecular thermochemistry, which is typically used to calculate reverse reaction rates from the relationship between the Gibbs free energy change of a reaction and the equilibrium constant.[3,4] The temperature-dependent Gibbs free energy of reaction can be computed from the enthalpies of formation, entropies, and heat capacities if one has means for accurately predicting those molecular properties.

In an ideal world, thermochemical properties for all relevant chemical species would be obtained from experiments or high-quality electronic structure calculations. Realistically, the cost associated with obtaining data for each species in this manner would be tremendous because the process of generating a large reaction mechanism may involve more than a million distinct species (including species in reactions later determined to be numerically negligible). An alternative method proposed several decades ago and still in use today is the group additivity method, which decomposes each molecule into groups and sums up the thermochemical contributions from each group. The group values were originally derived from a regression on experimental data,[5] but today most group values are derived from quantum chemistry calculations.[6–8] Group additivity can be applied very rapidly to large numbers of molecules and can provide highly accurate results for some classes of molecules. For example, the thermochemistry of hydrocarbons without cycles is predicted particularly well,[9] but more exotic species, especially heteroatom-containing and fused cyclic compounds, are ill-suited for the group additivity method. Careful collection of new data, manual selection of new groups, and a renewed fitting procedure are required every time incompatible species are encountered, although there have been some efforts toward automatic group selection for a limited set of molecules.[10]

Alternatives are provided by the much more flexible frameworks arising in the field of machine learning. A host of new machine learning methods, especially deep learning methods, have become available for not only the classical areas of computer vision and natural language processing[11] but also for chemical property predictions.[12–18] Machine learning methods can easily be applied to different chemical domains, and training a model that is useful across a broad range of chemistry does not require significant manual engineering of features. The downside is that most methods require very large molecular data sets for training, which are usually only available at low levels of theory.[19] In addition, machine learning models most often treat molecules as rigid structures or graphs, even though effects due to different conformers, especially for entropy, are important in reality.[20]

To overcome the limitation of data set size, a common technique in machine learning called transfer learning can be employed, in which knowledge learned by training in one domain is transferred to a second domain.[21] In this context, the first domain is a large quantity of low-level density functional theory (DFT) calculations and the second domain is a much smaller collection of thermochemical data from experiments and high-quality quantum mechanical calculations. The information gained from training on a wide array of chemistry greatly enhances the ability to learn from the limited amount of high-level data. Transfer learning and a related technique, Δ-machine learning, have already been successfully employed for energy predictions of molecular geometries,[22,23] and the benefit of transfer learning has been explored across many molecular data sets.[24]

Because high-accuracy data are scarce, our first goal is to construct an enthalpy of formation data set composed of high-quality explicitly correlated coupled cluster calculations. We also constructed entropy and heat capacity data sets using high-quality DFT calculations. Moreover, we wish to further improve the quality of the enthalpy data by deriving bond additivity corrections (BACs), which are a simple method to correct for systematic errors in energy calculations of electronic structure methods.[25] After the new data sets are supplemented with experimental data, the second goal is to train transfer learning models that leverage both existing large low-quality data sets and the newly created, but much smaller, high-quality data sets to obtain models that yield predictions of high accuracy. Furthermore, we aim to create an entropy prediction model capable of accounting for conformational effects by carefully selecting its training data. The ultimate goal is to use the models as part of the Reaction Mechanism Generator (RMG) software[3] and to replace its group additivity scheme.

3.2 Computational details

3.2.1 Transfer learning

Transfer learning is a frequently used technique in the machine learning community in which knowledge learned by a model on some task is applied to a different task. Often, a lot of data are available for simpler prediction tasks while only limited data exist in related domains of interest. In the context of molecular property prediction, obtaining large amounts of training data using low-level DFT calculations is a straightforward task, but compiling large sets of wave function-based calculations is associated with significantly higher cost. A transfer learning model in this realm is initially trained on low-level DFT calculations and subsequently refined using the limited high-accuracy data.

A schematic of the complete model used here is shown in Figure 3.1. The molecular representation and the foundation for the model are based on models used in the studies by Li et al.[26] and by Coley et al.,[18] both of which are based on so-called graph convolutions.[27,28] We refer the interested reader to the descriptions provided in those papers and will limit ourselves here to a concise explanation with a more in-depth treatment of the transfer learning module. Molecules are represented as labeled undirected graphs $M = (A, B)$, which are ordered pairs of vertices A corresponding to the atoms and edges B corresponding to the bonds. A bond is then given by the unordered pair of atoms $(x, y) \in B$ with $x, y \in A$. Each atom $a \in A$ is associated with an atom feature vector $\mathbf{f}_a$ which aggregates the following descriptors: atomic number, the number of non-hydrogen neighbors (heavy atoms), the number of hydrogen neighbors, and ring membership. Similarly, each bond $(a, y) \in B$ is associated with a bond feature vector $\mathbf{f}_{ay}$ only containing information about ring membership. Conventional models might use the bond order and aromaticity indicators as additional features, but these were not included here because the model was found to perform equally well without. This also removes the requirement of selecting specific resonance structures to train on. The ring membership descriptor counts how many rings of each size an atom or a bond is part of. Effectively, this encodes a kind of simplified representation of global molecular structure in the feature vector. All other features only describe the local neighborhood around an atom and the atom itself. Alternatively, or in addition to this, a global attention mechanism[29] could be added on top of the graph convolutions to incorporate distal information. The base model is composed of a graph convolutional neural network that converts the molecular representation described by the

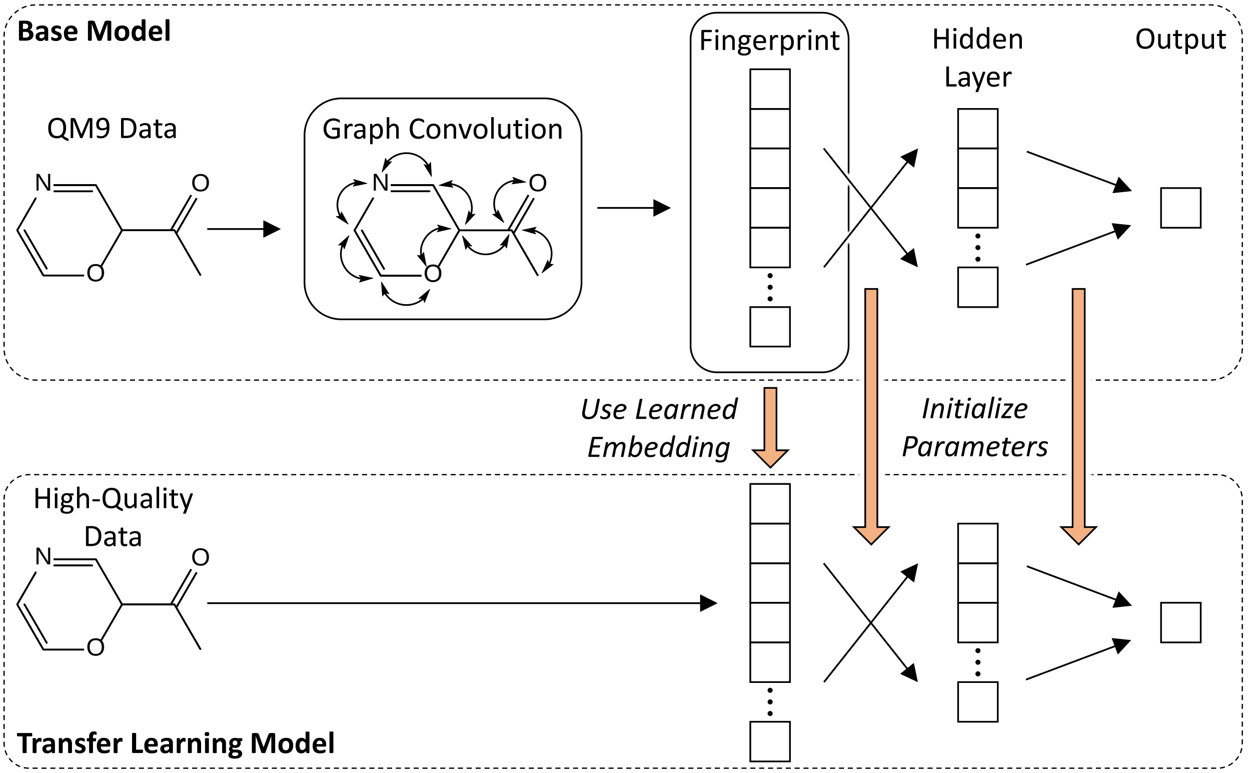


Figure 3.1. Transfer learning model architecture using a base model to learn a molecular embedding and neural network parameter initialization.

set of all $\mathbf{f}_a$ and $\mathbf{f}_{ay}$ to a fixed-length molecular feature vector (fingerprint), which is then passed through a final hidden layer before the output layer. The output vector has a single element for the enthalpy of formation model and for the entropy model, and seven elements for the heat capacity model in order to predict heat capacities at seven different temperatures simultaneously. The graph convolution essentially takes each $\mathbf{f}_a$ and $\mathbf{f}_{ay}$ and passes them through neural network layers to incorporate nearest neighbor features into new feature vectors for each atom. This process is repeated for a total of three iterations, thus incorporating information up to a depth of three into the feature vector for each atom. Subsequently, the resulting atom feature vectors are combined and sparsified using a *softmax* activation function to yield the molecular fingerprint.

The base models are trained on B3LYP/6-31G(2df,p) data. The transfer learning models are separate models trained on CCSD(T)-F12/cc-pVDZ-F12//B3LYP/6-31G(2df,p) data with bond additivity corrections for enthalpy of formation and on ωB97X-D3/def2-TZVP for entropy and heat capacities. The quantum mechanical data for the transfer learning models are combined with experimental data. Then, 5% of the available training data for each model were used as validation data sets for early stopping and separate test sets are used to measure model performance. A more detailed description of the data sets will follow in Section 3.3. The transfer learning models do not retrain the graph convolutions and instead use the learned fingerprint embedding from the base models directly. Additional knowledge is transferred from the base models to the transfer learning models by initializing their weights using the fully trained weights from the base models.

All models are trained using the *Adam* optimizer[30] with a mean squared error loss function and a batch size of one. The learning rate is given by $\eta \exp(-t/\tau)$ where η is the initial learning rate, t is the epoch, and τ is a decay time constant. The molecular fingerprints have a length of 512, the final hidden layer contains 50 units with a hyperbolic tangent activation function, and the output layer is linear. For all models, 5% of the training data were reserved for early stopping, and the learning rate parameters were set to $\eta = 7 \times 10^{-4}$ and $\tau = 30$. The hyperparameters vary for the transfer learning models. We considered two different enthalpy models: one trained on experimental and coupled cluster data and one trained on only coupled cluster data. For the former, $\eta = 2 \times 10^{-5}$, and for the latter, $\eta = 5 \times 10^{-5}$, and $\tau = 30$ for both. Entropy and heat capacity models used $\tau = 10^3$. The entropy model further used $\eta = 2 \times 10^{-4}$, whereas the heat capacity model used $\eta = 5 \times 10^{-5}$. The results were most sensitive to batch size and learning rate. Before training, the data for the entropy and heat capacity models were normalized by subtracting the mean and dividing by the standard deviation of the training data.

While all of the models were trained with a mean squared error loss function, a more intuitive metric for assessing results is the mean absolute error (MAE), which will be reported throughout this paper. In addition, 95% confidence intervals (CI) calculated as twice the root-mean-square error (RMSE), which are commonly reported in thermodynamic tables,[31] are listed as well. The models were trained using the *DataDrivenEstimator* package available on GitHub.[32]

3.2.2 Thermochemistry calculations

Electronic structure calculations were performed at a variety of levels of theory. The B3LYP/6-31G(2df,p) level of theory is used for low-level geometry optimizations and frequency calculations (used for calculating low-level enthalpies of formation, entropies, and heat capacities). A scale factor of 0.965 is applied to the computed harmonic frequencies used to compute the zero-point energy (ZPE).[33] High-level geometry optimizations and frequency calculations were performed at the ωB97X-D3/def2-TZVP level of theory corrected by a scale factor of 0.975 (used for calculating entropies and heat capacities at the higher level of theory).[34] High-level energies were calculated at the CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) level of theory (used for calculating high-level enthalpies of formation). The double-ζ basis set was selected in order to allow for a large number of coupled cluster calculations while maintaining reasonable accuracy (the accuracy after fitting bond additivity corrections is shown in Section 3.4.1). The geometries for some molecules selected for CCSD(T)-F12 calculations were taken directly from the published QM9 data set.[19] They were not reoptimized, and we did not attempt to confirm that they are the lowest-lying conformers. For new geometry calculations, the lowest energy conformer was selected based on a conformer search using RDKit[35] and the MMFF94 force field. For enthalpy calculations at 298 K, contributions from other conformers can mostly be neglected.[20] On the other hand, entropy is more strongly affected by conformational variations, so we only calculated molecules without rotatable bonds for the entropy models and for molecules with rotatable bonds we captured conformational effects implicitly by using experimental data (a more detailed description of the data sets is available in Section 3.3). All DFT calculations made use of the Q-Chem 5.1 electronic structure code[36] and the coupled cluster calculations used Molpro 2015.[37–39]

Standard ideal gas statistical thermodynamic models were used to compute rigid rotor harmonic oscillator (RRHO) partition functions. Enthalpies and entropies were calculated at 298 K and heat capacities were calculated at seven different temperatures—300 K, 400 K, 500 K, 600 K, 800 K, 1000 K and 1500 K. Symmetry contributions are not included in the entropies because RMG incorporates them automatically during mechanism generation. Therefore, the partition function for entropy was not divided by the external symmetry number. In fact, training a machine learning model to predict entropies that are symmetry-corrected is a more difficult task because the model has to implicitly learn symmetry numbers, which instead could easily be applied after training a model that does not include symmetry. Of course, correct computational determination of symmetry numbers, whether by estimating point groups from three-dimensional molecular geometries or from a computation based on a molecular graph representation, is a complex task in itself already discussed in the literature[40–42] and is outside the scope of this study.

Calculation of the enthalpy of formation follows the atomization energy approach detailed by Curtiss et al.[43] using the atomization energy supplemented with experimental data for the atoms. The enthalpy of formation at 298 K for molecule M is given by

$$\Delta_{\mathrm{f}} H^{\circ}_{298}(M) = \Delta_{\mathrm{f}} H^{\circ}_{0}(M) + \Delta H^{\circ}_{0\to 298}(M) - \sum_{a\in A(M)} \Delta H^{\circ}_{0\to 298}(\mathrm{element}_a) \tag{3.1}$$

where $\Delta_{\mathrm{f}} H^{\circ}_{0}(M)$ is the calculated enthalpy of formation at 0 K and the other two terms are corrections to go from 0 K to 298 K for the molecule and the atoms, respectively. $\Delta H^{\circ}_{0\to 298}(M)$ is the enthalpy increment for the molecule evaluated from ideal gas partition functions. $\Delta H^{\circ}_{0\to 298}(\mathrm{element})$ are the enthalpy corrections for atoms, which are known and tabulated, e.g., in CODATA.[44] It is an experimental correction for elements in their reference states renormalized per constituent atom. For example, the correction for carbon corresponds to graphite and the correction for hydrogen corresponds to one half of H_2(g). The enthalpy of formation at 0 K is given by

$$\Delta_{\mathrm{f}} H^{\circ}_{0}(M) = \left[\sum_{a\in A(M)} \Delta_{\mathrm{f}} H^{\circ}_{0}(a)\right] - \Delta_{\mathrm{at}} H^{\circ}_{0}(M) \tag{3.2}$$

where $\Delta_{\mathrm{f}} H^{\circ}_{0}(a)$ denotes the experimental enthalpy of formation at zero Kelvin for a gas phase

Table 3.1. Atomic energies, experimental enthalpies of formation, and enthalpy corrections in kcal mol^{-1}.

Element a	$E_0(a)^a$ (CCSD(T)-F12)	$E_0(a)^b$ (B3LYP)	$\Delta_\mathrm{f} H_0^\circ(a)^c$	$\Delta H^\circ_{0\to298}(\mathrm{element}_a)^c$
H	−313.64	−313.93	51.630 ± 0.001	1.01
C	−23 712.31	−23 749.21	169.98 ± 0.10	0.25
N	−34 215.55	−34 251.89	112.53 ± 0.02	1.04
O	−47 060.07	−47 103.74	58.99 ± 0.02	1.04

[a]Calculated using RCCSD(T)-F12a/cc-pVDZ-F12 with spin-orbit corrections from Ref. [47, 48]. [b]From Ref. [19]. [c]From Ref. [43].

atom a with reference to its standard state. These experimental enthalpies are for the individual atoms themselves, for example, $\Delta_\mathrm{f} H_0^\circ(\mathrm{C})$ corresponds to the carbon atom and not to graphite. Highly accurate values are available for some atoms in the Active Thermochemical Tables.[31,45] The atomization energy at 0 K is given by

$$\Delta_\mathrm{at} H_0^\circ(M) = \left[\sum_{a\in A(M)} E_0(a)\right] - E_0(M) \tag{3.3}$$

where $E_0(M)$ denotes the calculated electronic energy of the molecule including its scaled zero-point energy. The electronic energy of an atom $E_0(a)$ usually includes spin-orbit corrections.[46] The calculated electronic energies as well as the experimental enthalpies for the atoms and their elemental enthalpy corrections are listed in Table 3.1.

3.2.3 Bond additivity corrections

Atomization energies obtained from *ab initio* calculations are often not very accurate, because atoms and standard-state forms of some elements (e.g., graphite, $O_2(^3\Sigma_g^-)$) have significantly different electronic states than the closed-shell organic molecules studied here. To improve the accuracy of the formation enthalpy, it is common to use bond additivity corrections (BACs), which are empirical corrections to molecular energies and enthalpies of formation that use a few fitted parameters to correct for systematic errors in electronic structure calculations for some bond types. Fitting the parameters to a set of relatively few (tens or hundreds) low-uncertainty experimental data can significantly improve the error of calculations relative to experimental data and generalizes well beyond molecules in the reference data set because the corrections are specific to atoms and bonds rather than the molecule as a whole.

Two different types of BACs are found in literature. One possible implementation corrects for specific bond types (C−H, C−C, C−O, C=C, etc.) by adding a parameter for each type that adds a correction to the molecular energy based on how many times the bond occurs in a molecule.[49] In this case, the BAC scheme is given by

$$\Delta_{\mathrm{f}}^{\mathrm{BAC}} H_{298}^{\circ}(M) = \Delta_{\mathrm{f}} H_{298}^{\circ}(M) + \sum_{b \in B(M)} C(b) \tag{3.4}$$

where $C(b)$ is a parameter for bond type b. By limiting the corrections to only explicitly encoded bond types, resulting energies could be problematic for molecules that do not fit well into the set of molecules that can be created from the set of bond types, such as molecules that possess significant resonant character. Extending the corrections to transition states and ions could also be problematic. Instead, we choose a method that does not require labeling of bond types by using bond distances and atom identities instead, which are readily available from electronic structure calculations. This procedure is based on the method developed by Anantharaman and Melius.[25] We note that while Anantharaman and Melius refer to the method as "bond additivity correction", there are no parameters for specific bond types, only for atoms, but we adopt their nomenclature for consistency in the literature.

Bond additivity corrections to the computed enthalpy of formation, $\Delta_{\mathrm{f}} H_{298}^{\circ}$, are composed of molecular, atomic, and bond-wise corrections:

$$\Delta_{\mathrm{f}}^{\mathrm{BAC}} H_{298}^{\circ}(M) = \Delta_{\mathrm{f}} H_{298}^{\circ}(M) - E_{\mathrm{BAC}}(M) - \sum_{a \in A(M)} \alpha(a) - \sum_{b \in B(M)} E_{\mathrm{BAC}}(b) \tag{3.5}$$

$E_{\mathrm{BAC}}(M)$ is a molecular correction term given by

$$E_{\mathrm{BAC}}(M) = K \left(S(M) - \sum_{a \in A(M)} S(a) \right) \tag{3.6}$$

where $S(M)$ and $S(a)$ are spin quantum numbers for the molecule M and the atom a, respectively. The first sum in Equation (3.5) captures corrections due to all atoms in the molecule where α is a fitted parameter for each atom type. The second sum in Equation (3.5) contains corrections for each bond present in the molecule and can be further decomposed as

$$E_{\mathrm{BAC}}\left((x, y)\right) = \sqrt{\beta(x)\beta(y)} e^{-\xi R_{xy}} + \sum_{w \in N(x) \setminus y} [\gamma(w) + \gamma(x)] + \sum_{z \in N(y) \setminus x} [\gamma(z) + \gamma(y)] \tag{3.7}$$

where R_{xy} is the bond distance between atoms x and y, β and γ are additional parameters for each atom type, the set of atoms immediately adjacent to atom a is denoted by $N(a)$, and ξ is set to $3\,\text{Å}^{-1}$ based on the value used by the developers of the method.[25] The first term in Equation (3.7) has a negative exponential dependence on bond distance, which means the term will be more significant for shorter distances common in unsaturated bonds. The second and third terms sum over the neighbors of x and y, respectively, thus correcting for errors due to adjacent atoms. Note that each

Table 3.2. Bond additivity correction parameters (kcal mol^{-1}).

Element	α	β	γ
H	−0.9000	14.5704	0.1501
C	2.8338	0.0048	−0.0346
N	1.8315	0.0336	0.0118
O	1.1964	0.0099	0.0024

sum contains the parameter corresponding to the atom whose neighbors are being summed over. The method uses three parameters per element, α, β, γ, for a total of only twelve parameters for the four elements HCNO. Compared to the BACs in Equation (3.4), this is typically a smaller number of parameters. Here, we do not fit the molecular correction shown in Equation (3.6) because it was not found to provide any additional improvement and destroys the size consistency of the electronic structure calculation.[25]

We fit BAC parameters for the high-level coupled cluster calculations but do not consider them for the low-level DFT-derived enthalpies of formation used to train the base model because it is only used to obtain a good molecular representation and suitable initialization for the high-level model. To fit the parameters, we performed a regression with mean squared error loss function. The fitted parameters are shown in Table 3.2.

In order to facilitate deriving bond additivity corrections in the future, we have implemented a system to automatically execute quantum chemistry calculations for a data set of reference species and to fit corrections for the calculated data. This procedure is described in more detail in Section 3.6 and an example application for a DFT method is shown there.

3.3 Data sets

Training effective machine learning models is to a certain extent an exercise in data set selection and curation. As such, we are using an array of representative data sets from literature and proprietary sources and some created by ourselves. Many of the electronic structure calculations and geometries are either taken directly from the popular QM9 data set[19] with up to nine non-hydrogen atoms, or subsets of molecules are selected from the set of all molecules in QM9 in order to be calculated at a different level of theory. QM9 properties, which include the results of energy and harmonic vibrational frequency calculations, are available at the B3LYP/6-31G(2df,p) level of theory. Because we are only interested in HCNO-containing molecules and because diffuse functions were not included in the basis set, we removed all fluorine-containing molecules. We also removed the set of molecules that failed the consistency check described in the original publication, which involves converting force field, semi-empirical, and density functional theory (DFT) geometries to InChI strings and verifying that they are identical.[19] Other than the high-accuracy data for fitting bond additivity

Table 3.3. Enthalpy of formation ($\Delta_f H^\circ_{298}$), entropy (S°_{298}), and heat capacity (C_p) data sets.

Data set	Name	Property	Level(s)	Size
1	`bac_fit`	$\Delta_f H^\circ_{298}$	CC[a], expt.	147
2	`bac_test`	$\Delta_f H^\circ_{298}$	CC[a], expt.	412
3	`base`	$\Delta_f H^\circ_{298}$, S°_{298}, C_p	DFT-low[b]	~130k
4	`tf_h_1`	$\Delta_f H^\circ_{298}$	CC[a]	~10k
5	`tf_h_2`	$\Delta_f H^\circ_{298}$	expt.	~3k
6	`tf_s`	S°_{298}	DFT-high[c]+expt.	~3k
7	`tf_c`	C_p	DFT-high[c]+expt.	~2k
8	`tf_h_test`[d]	$\Delta_f H^\circ_{298}$	GA[e], DFT-low[b], CC[a]+expt.	~1.2k
9	`tf_s_test`[d]	S°_{298}	GA[e], DFT-low[b], DFT-high[c]+expt.	~0.3k
10	`tf_c_test`[d]	C_p	GA[e], DFT-low[b], DFT-high[c]+expt.	~0.2k

[a] CCSD(T)-F12/cc-pVDZ-F12//B3LYP/6-31G(2df,p) + BAC. [b] B3LYP/6-31G(2df,p). [c] ωB97X-D3/def2-TZVP. [d] Contain the same molecules. [e] Group additivity.

corrections, experimental data were obtained from a version of the NIST-TRC database,[50] henceforth simply referred to as NIST data. An overview of the data sets is given in Table 3.3. We only considered species with an even number of electrons, i.e., no doublet radicals.

For fitting the bond additivity corrections, we selected a data set of highly accurate experimental enthalpies of formation (`bac_fit`) and calculated the corresponding coupled cluster enthalpies of formation. The uncertainty in each experimental enthalpy value in the `bac_fit` data set is at most 0.5 kcal mol^{-1}, but the majority are significantly lower. In thermodynamic tables, uncertainty is typically provided as 95% confidence intervals, which approximately correspond to two standard deviations to the left and to the right of the mean.[31] For the most part, the uncertainty in these data adhere to that standard, but we were not able to verify the uncertainty quantification used in some of the sources. This set of 147 enthalpy values spans diverse chemical species of both small and large size involving most permutations of bonds between HCNO atoms. It is obtained from a variety of sources[49,51–53] including the Active Thermochemical Tables.[31,45] We selected an additional set of 412 molecules from the NIST data for testing the fitted BACs (`bac_test`). The test set molecules are selected to have a more varied set of molecules; in particular `bac_test` includes somewhat larger molecules and potentially more complex electronic structure effects. Unlike `bac_fit`, the uncertainties of `bac_test` are not readily available, so there is an assumption that the experimental data are reasonably well-known.

As described earlier, enthalpies of formation, entropies, and heat capacities are first trained on a large data set of low-quality data (`base`) and then on a smaller data set of high-quality data. The low-quality data (`base`) are taken as the roughly 129 000 HCNO molecules in the QM9 set filtering out those with identified inconsistencies, supplemented by 1700 molecules from the NIST-TRC, bac_fit, and bac_test data sets recalculated at the B3LYP/6-31G(2df,p) level of theory to match the level of QM9 (base). The additional molecules correspond to those that do not overlap with the species already present in QM9.

There are three different transfer learning models: an enthalpy of formation model, an entropy model, and a heat capacity model. For each of these models, the training data sets are composed of both calculated and experimental data. The experimental data contain many molecules that are significantly larger than those in QM9 with up to 42 non-hydrogen atoms.

For the enthalpy of formation model, the high-quality training data are a combination of the 147 experimental data points for fitting BACs (bac_fit), experimental data for about 2700 NIST molecules (tf_h_2), and a selection of approximately 9800 explicitly correlated coupled cluster calculations (CCSD(T)-F12a/cc-pVDZ-F12//B3LYP/6-31G(2df,p) + BAC) corresponding to molecules sampled at random from QM9 (tf_h_1). Note that nonrandom selections can improve model performance,[26,54] but such an active selection scheme is not the focus of the present study.

The entropy model is trained on 2300 NIST data and 900 ωB97X-D3/def2-TZVP DFT calculations (tf_s). The NIST entropy data are of mixed accuracy, with some data from direct experimental measurements but much of the data coming from indirect methods and extrapolations. Internal and external symmetry number contributions are calculated using RMG for each NIST molecule and are removed from the experimental entropy because the goal is to train a model that can be used in RMG, which adds its computed symmetry contributions in during a reaction mechanism simulation. The 900 DFT calculations correspond to molecules randomly selected from QM9 with the constraint of being exclusively composed of cyclic or polycyclic cores without rotatable bonds.

The heat capacity model is trained on 1100 NIST data points and the same 900 ωB97X-D3/def2-TZVP DFT molecules (tf_c). The experimental heat capacities are from a mix of direct and indirect methods, with considerable variance in error bars. Then, 5% of the training data available for each model were reserved as a held-out validation data set to stop the training before overfitting.

Lastly, we selected molecules for test sets (tf_h_test, tf_s_test, tf_c_test) that are not present in any of the training data sets. For each property, the selection consists of roughly 10% of all molecules with available high-accuracy data. Because all molecules for which high-accuracy data are available are also present in the low-accuracy training data, each molecule in the test sets has both high- and low-accuracy properties. For example, none of the molecules in tf_s_test are in tf_s or base.

3.4 Results and discussion

3.4.1 Bond additivity corrections

As outlined previously, we calculated enthalpies of formation using the atomization energy method and then added corrections. Here, we compare the accuracy of the calculated values with and without fitted bond additivity corrections (BACs). The level of theory for the single point energy calculations is CCSD(T)-F12a/cc-pVDZ-F12, which will also be referred to as F12 for simplicity. BACs are fitted to minimize the difference between F12 enthalpies of formation calculated from quantum chemistry and the corresponding experimental data for the `bac_fit` data set in Table 3.3.

Overall, the fitting procedure reduced the average error across `bac_fit` very significantly. Before adding BACs, the MAE between the F12 enthalpies of formation and the high-accuracy experimental data was 8.98 kcal mol^{-1} and the RMSE was 10.45 kcal mol^{-1}. This very large error is surprising considering it has previously been shown that chemical accuracy (1 kcal mol^{-1}) is possible with a double-ζ basis for certain molecular reaction energies.[55] Of course, there is a large error-canceling effect between reactant and products for both atoms and bonds when calculating reaction energies that does not occur for enthalpies of formation calculated from atomization energies because the electronic structure of molecules and atoms is very different. Figure 3.2 shows that the error across `bac_fit` increases roughly linearly with increasing numbers of heavy atoms in a molecule. Such a trend indicates there is a large systematic error in the uncorrected values that scales with the number of heavy atoms. Knizia et al. report an MAE of 1.86 kcal mol^{-1} for CCSD(T)-F12a with a double-ζ basis set for atomization energies,[55] which is significantly smaller than our error. However, they are benchmarking against conventional CCSD(T)/CBS instead of experimental data and 47 out of the 49 molecules in their benchmark data set only have one or two non-hydrogen (heavy) atoms each. As shown in Figure 3.2, The errors for molecules in the `bac_fit` data set containing one or two heavy atoms are in line with those reported by Knizia et al. on their test set. We are not aware of any discussion regarding abnormal enthalpies of formation with double-ζ CCSD(T)-F12a in the literature, potentially because most studies that employ explicitly correlated coupled cluster methods use triple-ζ and larger basis sets, which are prohibitively expensive for the present study.

After adding fitted BACs, the MAE computed across `bac_fit` is reduced to only 0.70 kcal mol^{-1} and the RMSE becomes 1.18 kcal mol^{-1} (95% CI: 2.36 kcal mol^{-1}). Furthermore, the systematic error in Figure 3.2 has been removed. In fitting BACs, the atomic energies $E_0(a)$ in Equation (3.3) have effectively been redefined by adding corrections for each element (given by the α values in Equation (3.5)). As hypothesized, the derived BACs generalize well, which is demonstrated by a reduction in MAE from 13.90 kcal mol^{-1} to 0.98 kcal mol^{-1} for the test set `bac_test` (RMSE decreases from 14.74 kcal mol^{-1} to 1.31 kcal mol^{-1}), demonstrating that double-ζ calculations can yield good results if corrected with BACs. The uncertainties in the experimental enthalpy of formation values for the test set are not known and are likely higher than the very accurate data in `bac_fit`,

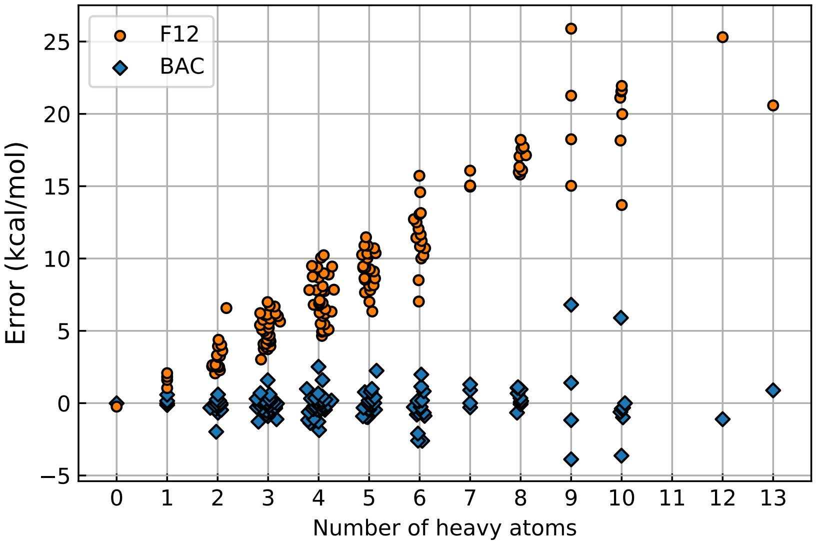


Figure 3.2. Enthalpy of formation errors vs. the number of heavy atoms in each molecule before (F12) and after (BAC) fitting bond additivity corrections on data set `bac_fit`.

which may be part of the reason for the slightly larger error across `bac_test`. Fitting parameters for specific bond types instead of using three parameters per atom type (using Equation (3.4) instead of Equation (3.5)) would lead to an almost identical reduction in error (MAE: 0.67 kcal mol^{-1}, RMSE: 1.16 kcal mol^{-1} for `bac_fit`) but may be sensitive to the resonance structure used for each molecule. Additionally, with that approach the atom corrections would be absorbed as part of the bond corrections instead of being treated separately.

The distribution of errors with and without BACs is shown in Figure 3.3. Each error is computed as the subtraction of the experimental value from the calculated value. The systematic error observed in Figure 3.2 manifests itself as a very wide range of errors much greater in magnitude than after fitting BACs. The majority of enthalpies of formation computed from atomization energies are in error by more than 5 kcal mol^{-1}. Including BACs leads to a tight distribution centered at zero with all but two molecules in error by less than 5 kcal mol^{-1}. The highest error of 6.80 kcal mol^{-1} corresponds to phenyl isocyanate. An error of such a large magnitude will cause issues in reaction mechanism generation, but the likelihood of such errors is small. Using a triple-ζ basis would most likely reduce the probability of large errors even further, but computational restrictions necessitated the use of a double-ζ basis here.

As before, the BAC procedure generalizes well to the test set `bac_test` as shown in the second panel of Figure 3.3. After fitting BACs, none of the molecules in `bac_test` are in error by more than 5 kcal mol^{-1}.

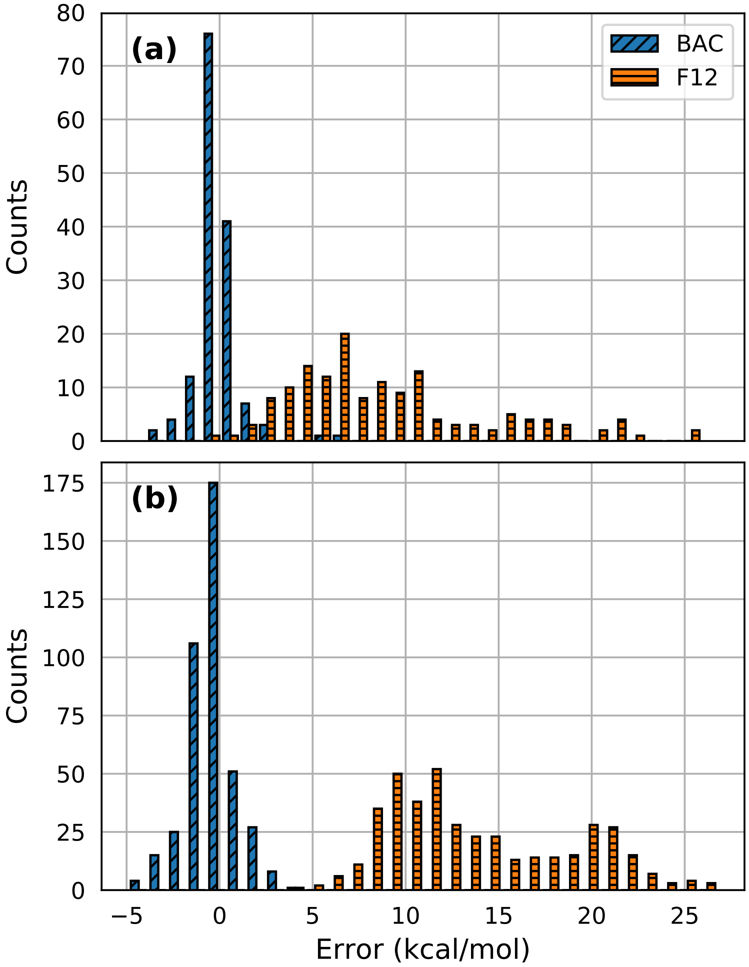


Figure 3.3. Distribution of enthalpy of formation errors relative to experiment (calculated minus experimental value) in the `bac_fit` data set **(a)** and the `bac_test` data set **(b)** before (F12) and after (BAC) fitting corrections.

3.4.2 Transfer learning

First, the three base models (Figure 3.1), one for enthalpy of formation, one for entropy, and one for heat capacities, were trained on the `base` data set (Table 3.3). In order to train the enthalpy of formation transfer learning model, BACs were applied to all high-level CCSD(T)-F12/cc-pVDZ-F12 data to form data sets `tf_h_1` and `tf_h_test`. The training data for the transfer learning enthalpy of formation model are the combination of the coupled cluster (`tf_h_1`) and experimental data (`tf_h_2`). Similarly, the training data for entropy and heat capacity are both a combination of high-level DFT and experimental data (`tf_s` and `tf_c`, respectively). The transfer learning models used the mapping that converts a molecular graph to a fixed-length vector learned during training of the base models. The remaining neural network parameters were initialized using the corresponding weights in the base models. For all models, the molecules in the test data sets, `tf_h_test`, `tf_s_test`, and `tf_c_test`, are identical and their properties are available at both the low and high level so that performance can be measured both in terms of precision and accuracy. In this context, model *precision* is measured by how well the model predictions match the values at the level of theory of the training data:

$$\text{precision} = \frac{1}{N} \sum_{i=1}^{N} \left| p_i^{\text{model}} - p_i^* \right| \tag{3.8}$$

Here, p_i^* is the value of the property at the same level of theory as the data used to train the model. Equation (3.8) corresponds to MAE and the equation for RMSE is analogous. Model *accuracy* is measured by how well the model predictions match the true values of the property, which are approximated using experimental data or coupled cluster data:

$$\text{accuracy} = \frac{1}{N} \sum_{i=1}^{N} \left| p_i^{\text{model}} - \hat{p}_i \right| \tag{3.9}$$

Here, $\hat{p}_i$ corresponds to the “true” value of the property. Naturally, high accuracy is the most desirable property of a machine learning model for property prediction. Additionally, RMG was used to calculate group additivity estimates of the thermochemical properties for the test set molecules to enable comparison to current RMG predictions.[3,8]

The *accuracies* are shown in Table 3.4 in terms of mean absolute error (MAE), root-mean-square error (RMSE), and 95% confidence interval (CI). For all three properties, the predictions afforded by the transfer learning model are clearly better than those of the base model and especially those of group additivity. Therefore, the transfer learning model is more suitable for simulations in RMG. Because of the molecules available in the QM9 database, the test set contains many complex structures, such as fused and bridged polycyclic compounds with several heteroatoms. These types of molecules are especially difficult to model with group additivity because contributions to thermochemistry are not solely additive across the groups present in the molecule but are strongly influenced by less local contributions like ring strain. Even though the group additivity scheme implemented in RMG has sophisticated ring strain corrections,[8] it lacks the ability to model many such molecules. If the test set were only composed of linear hydrocarbons, it would be very likely that group additivity would outperform the transfer learning model since group additivity was trained to even higher-accuracy data than most of the training data used here. For more complex RMG simulations involving fused cyclic molecules, the transfer learning model is a better choice.

The *precisions* for the base models are shown in Table 3.5. The precisions for the transfer learning models are identical to their accuracies, so they are the values in Table 3.4. Except for the MAE corresponding to enthalpy of formation, the precisions of the base models are significantly worse than those of the transfer learning models, which is surprising given the amount of data the base models were trained on. However, this effect is compounded by the fact that many of

Table 3.4. Test set (`tf_h_test`, `tf_s_test`, `tf_c_test`) *accuracies* of enthalpy of formation ($\Delta_\mathrm{f}H^\circ_{298}$), entropy ($S^\circ_{298}$), and heat capacity ($C_\mathrm{p}$) predictions for the transfer learning models (TF), the base models, and group additivity (GA). $\Delta_\mathrm{f}H^\circ_{298}$ in $\mathrm{kcal\,mol^{-1}}$ and $S^\circ_{298}/C_\mathrm{p}$ in $\mathrm{cal\,mol^{-1}\,K^{-1}}$.

$\Delta_\mathrm{f}H^\circ_{298}$	MAE	RMSE	95% CI
TF	1.78	2.80	5.60
Base	4.76	6.47	12.94
GA	9.99	16.17	32.35
S°_{298}	MAE	RMSE	95% CI
TF	0.80	1.16	2.31
Base	9.74	13.67	27.34
GA	11.25	18.51	37.02
C_p[a]	MAE	RMSE	95% CI
TF	0.74	1.21	2.41
Base	2.48	3.32	6.63
GA	3.41	5.44	10.88

[a]Average across seven temperatures.

Table 3.5. Test set (`tf_h_test`, `tf_s_test`, `tf_c_test`) *precisions* of enthalpy of formation ($\Delta_\mathrm{f}H^\circ_{298}$), entropy ($S^\circ_{298}$), and heat capacity ($C_\mathrm{p}$) predictions for the base models only. $\Delta_\mathrm{f}H^\circ_{298}$ in $\mathrm{kcal\,mol^{-1}}$ and $S^\circ_{298}/C_\mathrm{p}$ in $\mathrm{cal\,mol^{-1}\,K^{-1}}$.

Property	MAE	RMSE	95% CI
$\Delta_\mathrm{f}H^\circ_{298}$	1.69	3.51	7.03
S°_{298}	1.50	2.23	4.46
C_p[a]	2.34	5.85	11.70

[a]Average across seven temperatures.

the molecules in the test set are drawn from the data with experimentally available properties, which are proportionally underrepresented in the training data for the base models compared to molecules drawn from QM9. For example, the MAE calculated across the *validation* sets used for early stopping (which are randomly drawn from the training data) is $1.56\,\mathrm{kcal\,mol^{-1}}$ for the base model and $1.42\,\mathrm{kcal\,mol^{-1}}$ for the transfer learning model, which are similar in magnitude. Similarly, for the validation data sets the MAE is $0.85\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ and $0.76\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ for the entropy base and transfer learning models, respectively, and $0.60\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ and $0.71\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ for the heat capacity base and transfer learning models, respectively. Regardless, this indicates that less than a tenth of all available molecules have to be calculated at the high level or obtained from experiment in order to train a model that reaches the same precision as a low-level model trained on all available molecules.

As mentioned in previous sections, entropy is strongly affected by conformational effects.[20] Ex-

Table 3.6. Test set (`tf_s_test`) *accuracies* and *precisions* of entropy (S°_{298}) predictions in $\mathrm{cal\,mol^{-1}\,K^{-1}}$ for the transfer learning model (TF) and the base model split by molecules with and without internal rotors.

Accuracy	MAE	RMSE	95% CI
TF (no rotors)	0.72	1.10	2.19
TF (with rotors)	0.85	1.19	2.37
Base (no rotors)	1.44	1.75	3.51
Base (with rotors)	14.12	16.85	33.70
Precision	**MAE**	**RMSE**	**95% CI**
Base (no rotors)	0.96	1.33	2.65
Base (with rotors)	1.78	2.58	5.16

perimental data naturally include all the important conformers and internal rotors. However, the quantum chemistry calculations used here are for a single conformer, and they do not include corrections for internal rotation. Table 3.6 shows that combining experimental and quantum chemistry calculations into a single training set affords predictions of nearly identical quality for molecules with and without internal rotors. Because the training data for the base model are composed exclusively of static electronic structure calculations, its accuracy and precision for molecules with internal rotors is lowered significantly.

Parity plots and frequency distributions of the errors for the different transfer learning models are shown in Figure 3.4. The error distributions show that while most molecules are predicted well by the transfer learning model, several predictions are poor, albeit more accurate than the base model and group additivity on average. In theory, prediction quality can be improved by providing more information in the form of input atom and bond featurization, for example, by incorporating molecular geometry, but molecular representation in RMG is inherently graph-based and lacks geometrical information. Furthermore, thermochemistry may be strongly affected by different molecular geometries. Rapid estimation of molecular geometries may be possible with distance geometry based three-dimensional embedding and subsequent force field optimization as is available in cheminformatics packages like the RDKit,[35] but exhaustive conformer searches for the selection of the lowest energy conformation may be prohibitively expensive for large molecules in RMG and it is not clear when distance geometry-based approaches might fail.

Obtaining nearly 10 000 high-level data points for the enthalpy of formation model, as was done in this study, is already associated with large computational cost. Therefore, it is important to know how much data are really needed to obtain acceptable results. To assess this, we trained different models on various fractions of the approximately 9800 F12 data (`tf_h_1` in Table 3.3) and tested on all of the remaining data. For example, the smallest training set considered by us is composed of 81 molecules with the test error being reported on the remaining 9724 molecules. The results are

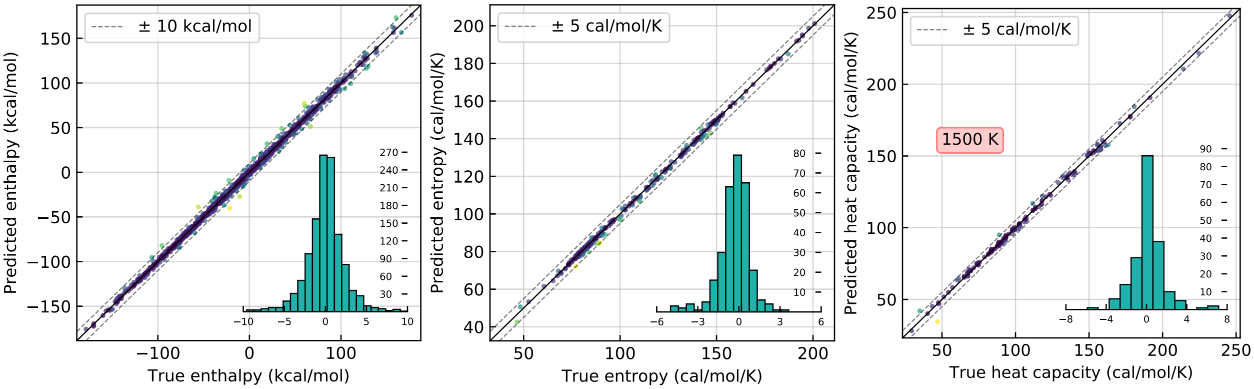


Figure 3.4. Parity plots of the experimental and high-level electronic structure calculations (“true”), and the values predicted by the transfer learning models (“predicted”) for the test sets `tf_h_test`, `tf_s_test`, and `tf_c_test`. Dashed lines of 10 kcal mol^{-1} and 5 cal mol^{-1} K^{-1} are shown to guide the eye. Frequency distributions of the signed errors (“predicted”−“true”) are superimposed. The heat capacity plot is for the values at 1500 K, which have the largest errors.

shown in Figure 3.5. Remarkably, the MAE is already smaller than 3 kcal mol^{-1} when only training on 81 molecules. This suggests that only very few data points are required to adapt the information learned during training of the base model to be suitable for predictions in the high-level domain. Predictions of practical importance can already be achieved with less than 1000 high-level training data, which is less than 1% of the low-level training data used in the base model. Interestingly, the lowest error in Figure 3.5 is smaller than the error in Table 3.4, even though the experimental molecules were not included here and the usual assumption in machine learning is that more data lead to smaller errors. However, the experimental data tend to be more diverse than the molecules in the `tf_h_1` data set, at least in terms of molecular size, which renders the learning task somewhat more difficult and may explain this difference.

3.5 Conclusions

With the continual development of new methods and the rapid expansion of molecular databases, machine learning is ideally suited for chemical property prediction in automated reaction mechanism generation. Because most methods are agnostic to the type of input molecule, machine learning frameworks are much more flexible than conventional (e.g., group additivity) ones. Nonetheless, the amount of required training data is usually very large, and these data are especially difficult to obtain at levels of theory that are of practical importance. To address this issue, we created an extensive data set of explicitly correlated coupled cluster enthalpies of formation, albeit still much smaller than available low-quality data sets. We fitted bond additivity corrections to reduce the mean absolute error compared to experiment to less than 1 kcal mol^{-1}. We also collected an array of experimental data and calculated additional high-level density functional data for new entropy and heat capacity data sets.

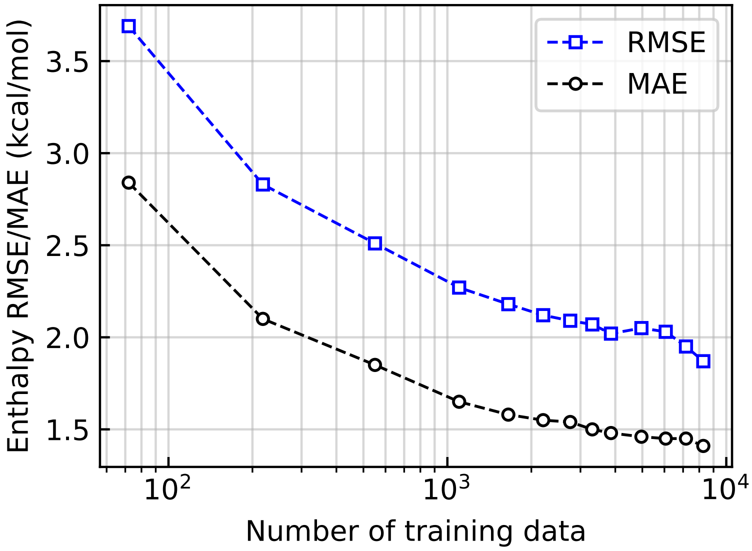


Figure 3.5. Test error of enthalpy of formation model with varying number of CCSD(T)-F12 training data points. The test error is computed across all molecules in data set `tf_h_1` that were not trained on.

In order to train useful machine learning models with the comparatively small amount of high-quality training data, we employed a transfer learning approach to predict enthalpy of formation, entropy, and heat capacities at several temperatures. Three base models were trained on 130 000 molecules, which were used to initialize parameters in the high-level neural network models and which provided learned molecular embeddings to convert molecular graphs into appropriate fixed-length vector representations. Subsequent training of the transfer learning models resulted in models capable of thermochemical property prediction with accuracies far exceeding those of the base models and group additivity. By combination of an experimental data set containing molecules with many rotatable bonds with a DFT data set only composed of rigid molecules, the entropy and heat capacity transfer learning models achieve equally accurate predictions for molecules with and without rotatable bonds. We showed that fewer than 1000 high-level training data points are required to obtain a useful enthalpy of formation model.

Several improvements can be made to both the methods and the data in the future. The larger error of the test sets compared to the validation data sets used for early stopping and the presence of predictions with large errors hint at issues with generalizability to significantly different chemical domains. To combat this issue, the current model design could be improved by incorporating novel ideas from methods that have been shown to generalize well to larger molecules, for example, incorporating 3D geometries into the graph convolution or constructing the convolution in a more atom-wise fashion.[13] The current data sets contain no radicals, but thermochemical predictions of radical species are still possible in RMG by using the hydrogen atom bond increment method (HBI) to predict their properties from their stable counterparts.[56] Alternatively, radicals could be directly included in the training data, thus directly enabling prediction of their properties without the need for HBI. Moreover, the developed models are limited to the realm of organic chemistry and extension to transition metal chemistry is not trivial due to difficulties with generating training data.

We have shown that transfer learning coupled with novel high-quality data are an effective technique to obtain accurate thermochemistry predictions suitable for automated reaction mechanism generation while only requiring small data sets on the order of a few thousand molecules. We expect that further refinement of the methods and data will lead to general-purpose property prediction schemes in the near future.

3.6 Appendix: Automated bond additivity corrections

In Section 3.2.3, we provide details on the two different types of bond additivity corrections and how they can be derived given a high-accuracy reference data set. To be useful for a wide range of users, many different BACs must be available. Different computational budgets and different molecular sizes necessitate different levels of theory for quantum chemistry calculations and, therefore, different BACs. In order to minimize the amount of manual effort required and to allow for straightforward comparisons between different methods, we developed an automated system to compute enthalpies of formation at the desired level of theory for a reference set of molecules and to derive the corresponding BACs.

3.6.1 Reference data and least squares regression

To be more generally useful, a more extensive and more accurate reference data set than that described in Section 3.3 is necessary. We curated a list of very low uncertainty data originating mostly from the *Active Thermochemical Tables*,[31,45] but some molecules were obtained from the *Third Millennium Ideal Gas and Condensed Phase Thermochemical Database for Combustion*,[57] from the *NIST Computational Chemistry Comparison and Benchmark Database*,[58] from Benson,[59] from the *NIST Chemistry WebBook*,[60,61] from Cioslowski et al.,[51] and from Pedley et al.[62] The vast majority of species have uncertainties that are at most $1\,\mathrm{kJ\,mol^{-1}}$ with the exception of some sulfur-containing molecules, for which up to $2.3\,\mathrm{kJ\,mol^{-1}}$ uncertainty was allowed because such data are difficult to obtain. The reference data include singlets, doublets, triplets, anions, cations, and are composed of H, C, N, O, S, F, Cl, and Br elements.

As mentioned in Section 3.2.3, two different BAC methods are popular: The method by Petersson et al.[49] (Petersson-type, Equation (3.4)) which fits one parameter per bond type and the method by Anantharaman and Melius[25] (Melius-type, Equation (3.5)) which fits three parameters per atom type and a molecular parameter. Both of these methods have been implemented in the Arkane thermodynamics and kinetics package available within RMG.[3] Fitting Petersson-type BACs is possible with a simple linear least squares regression, whereas the nonlinear term in Equation (3.7) necessitates a nonlinear least squares solution for Melius-type BACs. We use a trust region reflective algorithm to solve the optimization.[63]

3.6.2 Test case: ωB97M-V/def2-TZVPD bond additivity corrections

In addition to the results reported in Section 3.4.1, we selected the ωB97M-V/def2-TZVPD DFT method as an additional test case, because it has been shown to be an extremely promising density functional.[64,65] If adding BACs to DFT calculations were to enable thermochemical predictions close to chemical accuracy ($1\,\mathrm{kcal\,mol^{-1}}$), then more expensive methods (e.g., coupled cluster) may not be necessary for the accuracy required in reaction mechanism generation and good thermochemical data could be obtained for molecules that are too large for more expensive methods.

The quantum chemical calculations and subsequent thermochemistry calculations were performed automatically for the molecules in the reference database at the ωB97M-V/def2-TZVPD level using the ARC software package.[66] The results after fitting are shown in Figure 3.6. As expected, even the accuracy before fitting was already excellent for a DFT method. After fitting, Petersson-type BACs slightly outperform Melius-type BACs in terms of mean errors but the results are similar. Indeed, these errors are comparable to some higher level coupled cluster methods and enable generating thermochemistry data of sufficient accuracy for reaction mechanism generation at DFT cost.

The results in Figure 3.6 are measured across the entire reference database. We also wanted to determine how the different types of BACs perform on a separate molecular test set. Figure 3.7 shows such a comparison. The test molecules were selected because they include molecules with nonstandard bond lengths, that is, those that might benefit from the explicit dependence on bond length in the Melius-type BACs (Equation (3.7)). For example, dinitrogen trioxide has an extremely

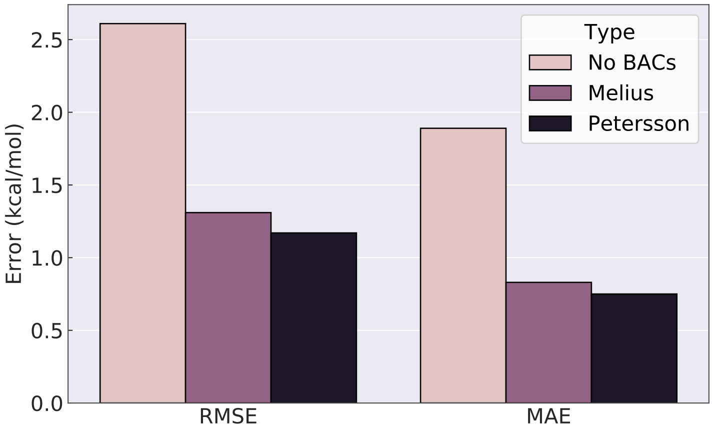


Figure 3.6. Comparison of Melius- and Petersson-type bond additivity corrections for the ωB97M-V/def2-TZVPD method.

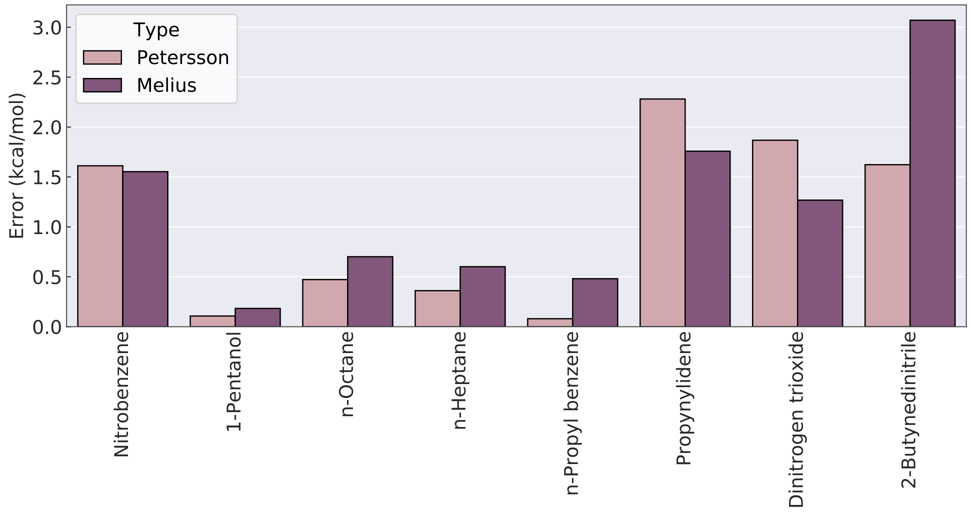


Figure 3.7. Comparison of Melius and Petersson corrections for a held-out test set of molecules, some of which contain unusual bond lengths. The molecules are sorted based on the magnitude of the difference of errors.

long N−N bond, whereas 2-butynedinitrile has short C≡N bonds. As shown in Figure 3.7, Melius-type BACs indeed perform better for unusually elongated bonds (e.g., nitrobenzene, dinitrogen trioxide) and molecules with nonstandard electronic structure (e.g., propynylidene), but they perform slightly worse for "normal" molecules and, surprisingly, significantly worse for the C≡N triple bonds in 2-butynedinitrile. Evidently, an explicit correction for the C≡N bond type in Petersson-type BACs provides the improved performance, whereas Petersson-type BACs cannot distinguish between a normal-length N−N single bond and a strongly elongated N−N single bond. Therefore, the choice of BAC-type will depend on the molecule that corrections are being applied to, although Petersson-type BACs seem to perform well in a wide variety of situations.

Figure 3.8 shows the parameter correlation matrix for Petersson-type BACs. Clearly, some parameters are strongly correlated. For example C(C−H) is strongly correlated with C(C−C), C(C−N), C(C−O), and C(C−S). This is not surprising because C−H generally appears with other single bonds such that their effect cannot be decoupled. The C(N−O) and C(N=O) parameters are also strongly correlated because the two bond types very often occur together in nitro groups. Therefore, this correlation is a result of labeling the nitro group resonance structure in a specific manner. Adding an additional parameter for bonds with non-integer bond order may reduce such effects.

The parameter correlation matrix for Melius-type BACs is shown in Figure 3.9. Even though there are fewer parameters, more parameters are correlated with each other. Most notably, the correlations between α and β parameters indicate that fitting atomic energy corrections in addition to bond corrections cannot generally occur in an uncorrelated fashion. Additionally, the strong

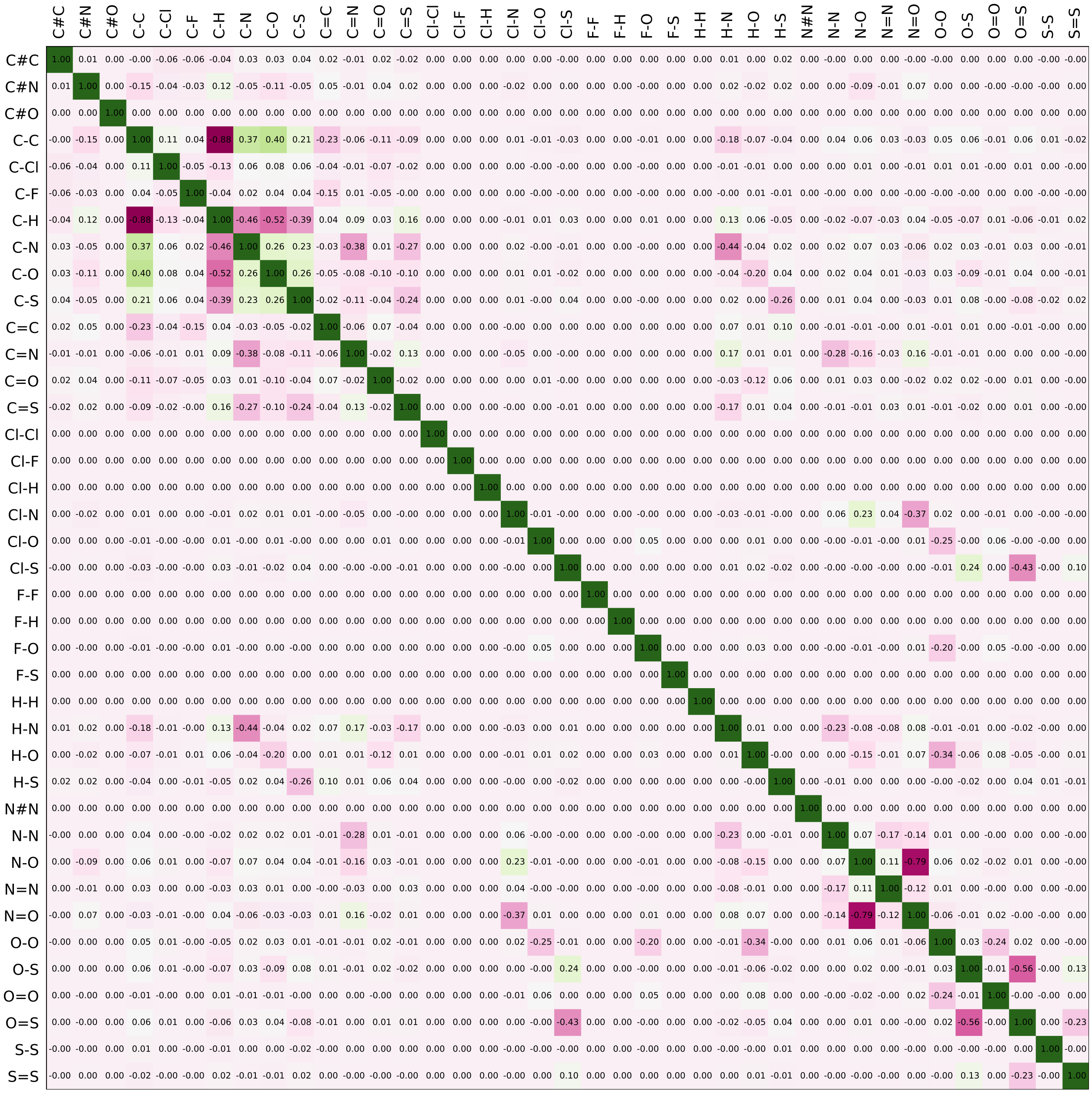
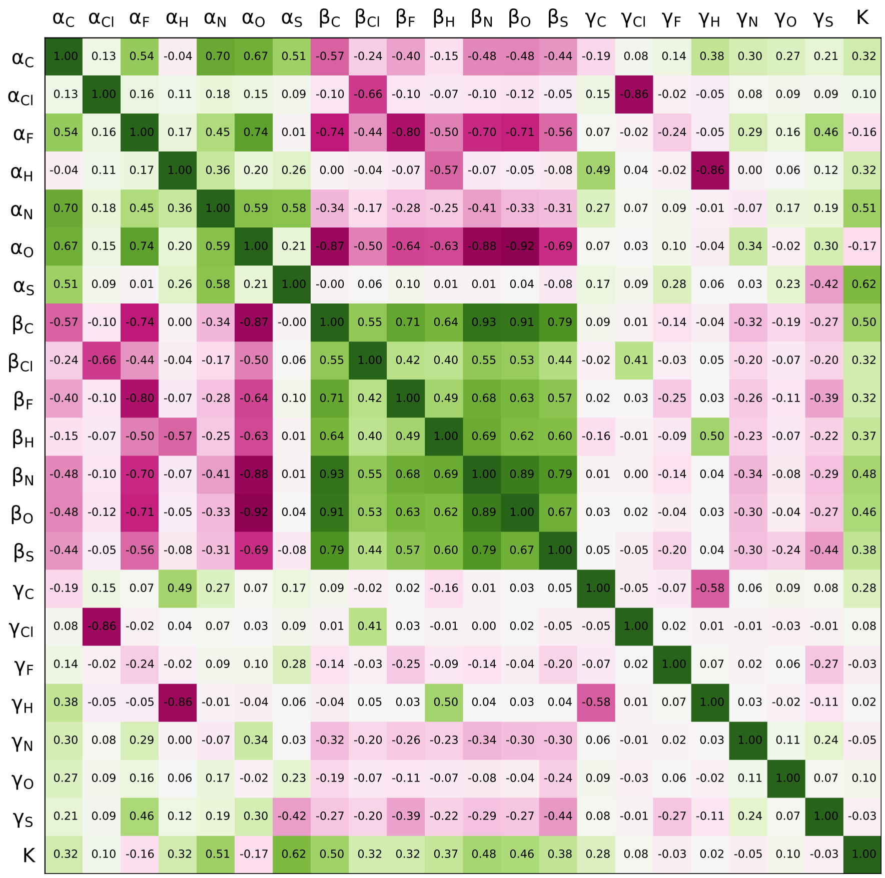


Figure 3.8. Parameter correlation matrix for Petersson-type bond additivity corrections.

correlations between different β parameters indicate that joining two β parameters into a single bond-type parameter like in Petersson-type BACs might be preferred since β parameters occur in pairs in Equation (3.7) anyway.

An issue with the reference dataset is that it is heavily biased towards common molecules, i.e., those containing many C−C and C−H bonds (e.g., alkanes). However, often users are interested in applying to BACs to less common molecules with nonstandard bonding and electronic environments, as it is more likely that calculations are not yet available for such molecules. In these cases, it would be desirable to more heavily weight the parameter estimation toward molecules with uncommon



	α_C	α_{Cl}	α_F	α_H	α_N	α_O	α_S	β_C	β_{Cl}	β_F	β_H	β_N	β_O	β_S	γ_C	γ_{Cl}	γ_F	γ_H	γ_N	γ_O	γ_S	K
α_C	1.00	0.13	0.54	-0.04	0.70	0.67	0.51	-0.57	-0.24	-0.40	-0.15	-0.48	-0.48	-0.44	-0.19	0.08	0.14	0.38	0.30	0.27	0.21	0.32
α_{Cl}	0.13	1.00	0.16	0.11	0.18	0.15	0.09	-0.10	-0.66	-0.10	-0.07	-0.10	-0.12	-0.05	0.15	-0.86	-0.02	-0.05	0.08	0.09	0.09	0.10
α_F	0.54	0.16	1.00	0.17	0.45	0.74	0.01	-0.74	-0.44	-0.80	-0.50	-0.70	-0.71	-0.56	0.07	-0.02	-0.24	-0.05	0.29	0.16	0.46	-0.16
α_H	-0.04	0.11	0.17	1.00	0.36	0.20	0.26	0.00	-0.04	-0.07	-0.57	-0.07	-0.05	-0.08	0.49	0.04	-0.02	-0.86	0.00	0.06	0.12	0.32
α_N	0.70	0.18	0.45	0.36	1.00	0.59	0.58	-0.34	-0.17	-0.28	-0.25	-0.41	-0.33	-0.31	0.27	0.07	0.09	-0.01	-0.07	0.17	0.19	0.51
α_O	0.67	0.15	0.74	0.20	0.59	1.00	0.21	-0.87	-0.50	-0.64	-0.63	-0.88	-0.92	-0.69	0.07	0.03	0.10	-0.04	0.34	-0.02	0.30	-0.17
α_S	0.51	0.09	0.01	0.26	0.58	0.21	1.00	-0.00	0.06	0.10	0.01	0.01	0.04	-0.08	0.17	0.09	0.28	0.06	0.03	0.23	-0.42	0.62
β_C	-0.57	-0.10	-0.74	0.00	-0.34	-0.87	-0.00	1.00	0.55	0.71	0.64	0.93	0.91	0.79	0.09	0.01	-0.14	-0.04	-0.32	-0.19	-0.27	0.50
β_{Cl}	-0.24	-0.66	-0.44	-0.04	-0.17	-0.50	0.06	0.55	1.00	0.42	0.40	0.55	0.53	0.44	-0.02	0.41	-0.03	0.05	-0.20	-0.07	-0.20	0.32
β_F	-0.40	-0.10	-0.80	-0.07	-0.28	-0.64	0.10	0.71	0.42	1.00	0.49	0.68	0.63	0.57	0.02	0.03	-0.25	0.03	-0.26	-0.11	-0.39	0.32
β_H	-0.15	-0.07	-0.50	-0.57	-0.25	-0.63	0.01	0.64	0.40	0.49	1.00	0.69	0.62	0.60	-0.16	-0.01	-0.09	0.50	-0.23	-0.07	-0.22	0.37
β_N	-0.48	-0.10	-0.70	-0.07	-0.41	-0.88	0.01	0.93	0.55	0.68	0.69	1.00	0.89	0.79	0.01	0.00	-0.14	0.04	-0.34	-0.08	-0.29	0.48
β_O	-0.48	-0.12	-0.71	-0.05	-0.33	-0.92	0.04	0.91	0.53	0.63	0.62	0.89	1.00	0.67	0.03	0.02	-0.04	0.03	-0.30	-0.04	-0.27	0.46
β_S	-0.44	-0.05	-0.56	-0.08	-0.31	-0.69	-0.08	0.79	0.44	0.57	0.60	0.79	0.67	1.00	0.05	-0.05	-0.20	0.04	-0.30	-0.24	-0.44	0.38
γ_C	-0.19	0.15	0.07	0.49	0.27	0.07	0.17	0.09	-0.02	0.02	-0.16	0.01	0.03	0.05	1.00	-0.05	-0.07	-0.58	0.06	0.09	0.08	0.28
γ_{Cl}	0.08	-0.86	-0.02	0.04	0.07	0.03	0.09	0.01	0.41	0.03	-0.01	0.00	0.02	-0.05	-0.05	1.00	0.02	0.01	-0.01	-0.03	-0.01	0.08
γ_F	0.14	-0.02	-0.24	-0.02	0.09	0.10	0.28	-0.14	-0.03	-0.25	-0.09	-0.14	-0.04	-0.20	-0.07	0.02	1.00	0.07	0.02	0.06	-0.27	-0.03
γ_H	0.38	-0.05	-0.05	-0.86	-0.01	-0.04	0.06	-0.04	0.05	0.03	0.50	0.04	0.03	0.04	-0.58	0.01	0.07	1.00	0.03	-0.02	-0.11	0.02
γ_N	0.30	0.08	0.29	0.00	-0.07	0.34	0.03	-0.32	-0.20	-0.26	-0.23	-0.34	-0.30	-0.30	0.06	-0.01	0.02	0.03	1.00	0.11	0.24	-0.05
γ_O	0.27	0.09	0.16	0.06	0.17	-0.02	0.23	-0.19	-0.07	-0.11	-0.07	-0.08	-0.04	-0.24	0.09	-0.03	0.06	-0.02	0.11	1.00	0.07	0.10
γ_S	0.21	0.09	0.46	0.12	0.19	0.30	-0.42	-0.27	-0.20	-0.39	-0.22	-0.29	-0.27	-0.44	0.08	-0.01	-0.27	-0.11	0.24	0.07	1.00	-0.03
K	0.32	0.10	-0.16	0.32	0.51	-0.17	0.62	0.50	0.32	0.32	0.37	0.48	0.46	0.38	0.28	0.08	-0.03	0.02	-0.05	0.10	-0.03	1.00

Figure 3.9. Parameter correlation matrix for Melius-type bond additivity corrections.

substructures that are underrepresented in the data. The code implemented in Arkane can perform this weighting automatically by first extracting substructures, all of which are shown in Figure 3.10, and computing the weights for each molecule in the least squares optimization so that molecules with underrepresented substructures obtain larger weights. The weight for each molecule M is given by

$$w_M = \frac{1}{|S_M|} \sum_{s \in S_M} \frac{1}{n_s} \tag{3.10}$$

where S_M is the set of substructures in the molecule and n_s is the total number of substructures of type s across all molecules in the reference database. Figure 3.11 shows a comparison of weighted and non-weighted Petersson fits for a test set of molecules that were selected based on the occurrence of "unusual" features. For example, sulfuryl chloride contains two atom types that are not common in many combustion and pyrolysis reaction mechanisms, and 1,3-benzodithiole-2-thione contains an unusual aromatic ring system involving multiple sulfur atoms. Figure 3.11 shows that substructure-weighting improves the performance for such molecules (e.g., 1H-imidazole, 1,3-benzodithiole-2-thione, 1,3,5-trioxane, dinitrogen trioxide, and sulfuryl chloride) while not resulting in unacceptable errors for more common molecules (e.g., octane, propyl benzene, pentanol).

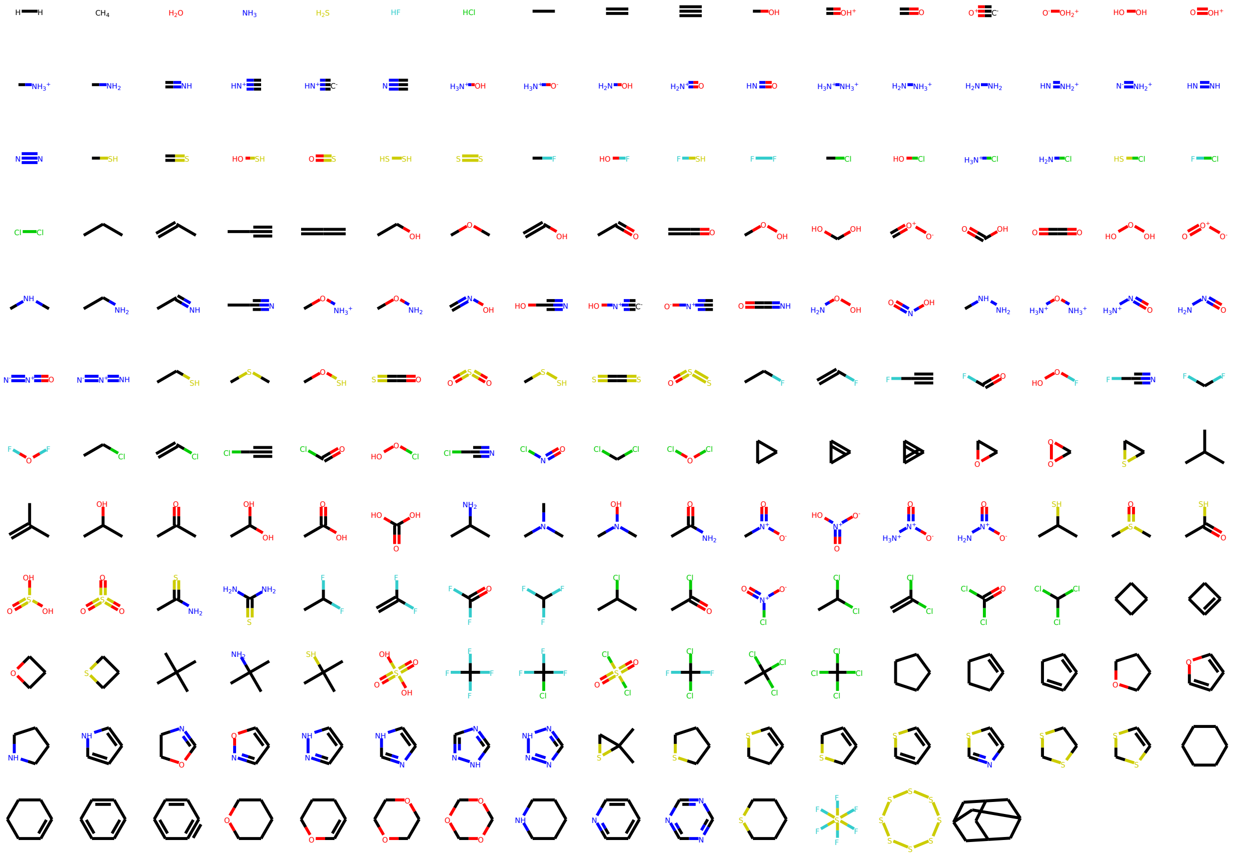


Figure 3.10. Substructures present in the bond additivity correction reference data.

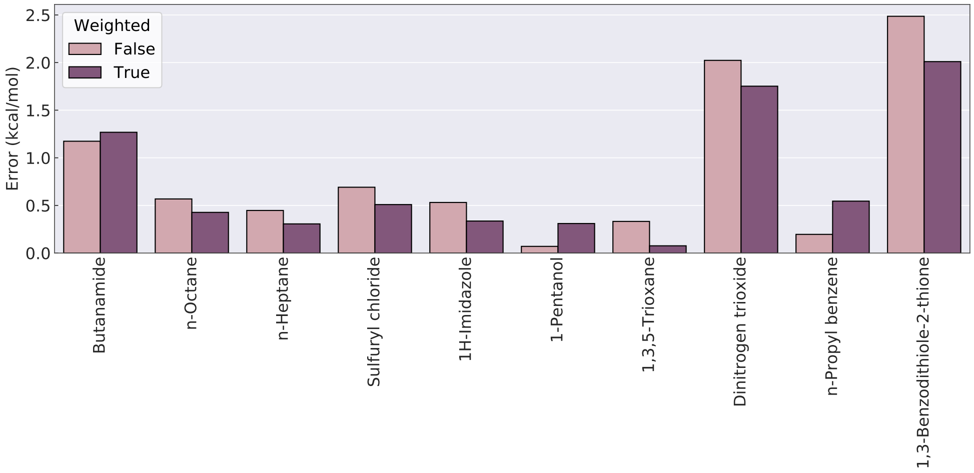


Figure 3.11. Comparison of unweighted and substructure-weighted Petersson fits for a held-out test set of molecules, containing "usual" (e.g., octane) and "unusual" (e.g., sulfuryl chloride) molecules alike. The molecules are sorted based on the magnitude of the difference of errors.

3.7 References

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

(2) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.

(3) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(4) Atkins, P.; de Paula, J., *Elements of Physical Chemistry*, 7th edition; Oxford University Press: New York, 2017.

(5) Benson, S. W., *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*, 2nd edition; Wiley: New York, 1976.

(6) Sumathi, R.; Green, W. H. Thermodynamic Properties of Ketenes: Group Additivity Values from Quantum Chemical Calculations. *J. Phys. Chem. A* **2002**, *106*, 7937–7949.

(7) Sebbar, N.; Bozzelli, J. W.; Bockhorn, H. Thermochemical Properties, Rotation Barriers, Bond Energies, and Group Additivity for Vinyl, Phenyl, Ethynyl, and Allyl Peroxides. *J. Phys. Chem. A* **2004**, *108*, 8353–8366.

(8) Han, K.; Jamal, A.; Grambow, C.; Buras, Z.; Green, W. An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation. *Int. J. Chem. Kinet.* **2018**, *50*, 294–303.

(9) Cohen, N.; Benson, S. W. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419–2438.

(10) He, T.; Li, S.; Chi, Y.; Zhang, H.-B.; Wang, Z.; Yang, B.; He, X.; You, X. An Adaptive Distance-Based Group Contribution Method for Thermodynamic Property Prediction. *Phys. Chem. Chem. Phys.* **2016**, *18*, 23822–23830.

(11) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.

(12) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**, arXiv: 1704.01212.

(13) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(14) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. **2017**, arXiv: 1706.08566.

(15) Hy, T. S.; Trivedi, S.; Pan, H.; Anderson, B. M.; Kondor, R. Predicting Molecular Properties with Covariant Compositional Networks. *J. Chem. Phys.* **2018**, *148*, 241745.

(16) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(17) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(18) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

(19) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

(20) Li, Y.-P.; Bell, A. T.; Head-Gordon, M. Thermodynamics of Anharmonic Systems: Uncoupled Mode Approximations for Molecules. *J. Chem. Theory Comput.* **2016**, *12*, 2861–2870.

(21) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.

(22) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

(23) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(24) Fare, C.; Turcani, L.; Pyzer-Knapp, E. O. Powerful, Transferable Representations for Molecules Through Intelligent Task Selection in Deep Multitask Networks. **2018**, arXiv: 1809.06334.

(25) Anantharaman, B.; Melius, C. F. Bond Additivity Corrections for G3B3 and G3MP2B3 Quantum Chemistry Methods. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.

(26) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(27) Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. **2015**, arXiv: 1509.09292.

(28) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

(29) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. **2014**, arXiv: 1409.0473.

(30) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**, arXiv: 1412.6980.

(31) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *Int. J. Quantum Chem.*, *114*, 1097–1101.

(32) Han, K.; Li, Y.-P.; Grambow, C. A. DataDrivenEstimator: A Package of Data Driven Estimators for Thermochemistry and Kinetics. https://github.com/ReactionMechanismGenerator/DataDrivenEstimator (accessed 01/07/2019).

(33) National Institute of Standards and Technology (NIST). Precomputed Vibrational Scaling Factors. https://cccbdb.nist.gov/vibscalejust.asp (accessed 10/19/2018).

(34) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

(35) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed 08/06/2018).

(36) Shao, Y. et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(37) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: A General-Purpose Quantum Chemistry Program Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 242–253.

(38) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M. *MOLPRO*, Version 2015.1, A Package of Ab Initio Programs, Molpro: Cardiff, U.K., 2015, 2015.

(39) Adler, T. B.; Knizia, G.; Werner, H.-J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127*, 221106.

(40) Ivanov, J.; Schüürmann, G. Simple Algorithms for Determining the Molecular Symmetry. *J. Chem. Inf. Model* **1999**, *39*, 728–737.

(41) Chen, W.; Huang, J.; Gilson, M. K. Identification of Symmetries in Molecules and Complexes. *J. Chem. Inf. Model* **2004**, *44*, 1301–1313.

(42) Vandewiele, N. M.; Van de Vijver, R.; Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B. Symmetry Calculation for Molecules and Transition States. *J. Comput. Chem.* **2015**, *36*, 181–192.

(43) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(44) Cox, J. D.; Wagman, D. D.; Medvedev, V. A., *CODATA Key Values for Thermodynamics*; Hemisphere: New York, 1989.

(45) Ruscic, B.; Bross, D. H. Active Thermochemical Tables (ATcT) Values Based on Ver. 1.122d of the Thermochemical Network. https://atct.anl.gov (accessed 10/10/2018).

(46) Ruscic, B.; Bross, D. H., Thermochemistry In *Mathematical Modelling of Gas-Phase Complex Reaction Systems: Pyrolysis and Combustion*, Faravelli, T., Manenti, F., Ranzi, E., Eds.; Computer-Aided Chemical Engineering, Vol. 45; Elsevier: Cambridge, MA, 2019.

(47) Huber, K. P.; Herzberg, G., *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*; Van Nostrand Reinhold Company: New York, 1979.

(48) National Institute of Standards and Technology (NIST). Electronic Spin Splitting Corrections. https://cccbdb.nist.gov/elecspin.asp (accessed 03/22/2019).

(49) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A.; Frisch, M. J. Calibration and Comparison of the Gaussian-2, Complete Basis Set, and Density Functional Methods for Computational Thermochemistry. *J. Chem. Phys.* **1998**, *109*, 10570–10579.

(50) NIST Thermodynamics Research Center. NIST/TRC Table Database, CD-ROM, 2004.

(51) Cioslowski, J.; Schimeczek, M.; Liu, G.; Stoyanov, V. A Set of Standard Enthalpies of Formation for Benchmarking, Calibration, and Parametrization of Electronic Structure Methods. *J. Chem. Phys.* **2000**, *113*, 9377–9389.

(52) Emel'yanenko, V. N.; Verevkin, S. P.; Varfolomeev, M. A.; Turovtsev, V. V.; Orlov, Y. D. Thermochemical Properties of Formamide Revisited: New Experiment and Quantum Mechanical Calculations. *J. Chem. Eng. Data* **2011**, *56*, 4183–4187.

(53) Månsson, M. Non-Bonded Oxygen-Oxygen Interactions in 2,4,10-Trioxaadamantane and 1,3,5,6,9-Pentoxecane. *Acta Chem. Scand. B* **1974**, *28*, 895–899.

(54) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(55) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 Methods: Theory and Benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.

(56) Lay, T. H.; Bozzelli, J. W.; Dean, A. M.; Ritter, E. R. Hydrogen Atom Bond Increments for Calculation of Thermodynamic Properties of Hydrocarbon Radical Species. *J. Phys. Chem.* **1995**, *99*, 14514–14527.

(57) Burcat, A.; Ruscic, B. *Third Millennium Ideal Gas and Condensed Phase Thermochemical Database for Combustion with Updates from Active Thermochemical Tables*; ANL-05/20 and TAE 960; Faculty of Aerospace Engineering, Technion – Israel Institute of Technology and Chemistry Division, Argonne National Laboratory, 2005.

(58) NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 20, https://cccbdb.nist.gov/, 2019.

(59) Benson, S. W. Thermochemistry and Kinetics of Sulfur-Containing Molecules and Radicals. *Chem. Rev.* **1978**, *78*, 23–35.

(60) Afeefy, H. Y. and Liebman, J. F. and Stein, S. E., Neutral Thermochemical Data In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Linstrom, P. J., Mallard, W. G., Eds., Gaithersburg, MD.

(61) Burgess, Jr., D. R., Thermochemical Data In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Linstrom, P. J., Mallard, W. G., Eds., Gaithersburg, MD.

(62) Pedley, J. B.; Naylor, R. D.; Kirby, S. P., Appendix: The CATCH Search and Retrieval System In *Thermochemical Data of Organic Compounds*, 2nd edition; Chapman and Hall: London, U.K., 1986.

(63) Branch, M. A.; Coleman, T. F.; Li, Y. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.* **1999**, *21*, 1–23.

(64) Mardirossian, N.; Head-Gordon, M. ωB97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(65) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(66) Dana, A. G.; Ranasinghe, D.; Wu, O. H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. H. ReactionMechanismGenerator/ARC: ARC 1.1.0, version 1.1.0, Zenodo. https://doi.org/10.5281/zenodo.3356849, 2019.

Chapter 4

Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry

Much of this work has previously appeared as Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**. Lagnajit Pattanaik performed the reaction template analysis. The code for generating new data is available at `https://github.com/cgrambow/ard_gsm`.

4.1 Background and summary

Rapid advancements in computational methods for chemical synthesis planning and automated reaction mechanism generation, especially in the area of machine learning, are causing a significant shift in how such problems are tackled. Deep learning approaches are replacing conventional quantitative structure-activity relationships often based on support vector machines, decision trees, or linear methods like partial least squares.[1,2] These new systems are becoming widely available for computer-aided retrosynthesis,[3] reaction outcome prediction,[3] high-throughput virtual screening,[4] and more general molecular property prediction.[5,6] Computational approaches are also increasingly common in reaction mechanism generation due to the large number of species and reactions that are generally required for accurate descriptions of phenomena like pyrolysis, combustion, and atmospheric oxidation.[7–9] Frequently, this involves characterizing chemical pathways with quantum chemistry,[8] but deep learning methods have also recently been applied to estimate thermochemistry during mechanism generation.[10,11]

While computers already outperform humans at qualitatively predicting reaction products[12,13] and successful yield predictions have been demonstrated for limited datasets,[14,15] quantitative reaction information is still elusive in large databases like Reaxys,[16] Pistachio,[17] and the United States

Patent and Trademark Office database.[18] Reaction yield, time, and some quantitative conditions like temperature are sometimes available, but there is usually no information on reaction kinetics. If such data were available, calculation of derived properties—such as minimum reaction times and branching ratios—would be possible. Our goal is to provide a quantitative dataset of reactions that enables the calculation of such data and can lead to more efficient drug design and help in deciding which reactions are important in mechanism generation.

Computationally generating a dataset of reactions is significantly more complex than only calculating stable equilibrium structures because transition states (TSs) of chemical reactions cannot be enumerated in the same manner as stable molecules. Even if the reactant and product structures are known, the exact TS geometry has to be found via a human-guided search or with expensive automated TS finding methods. Here, we use automated potential energy surface exploration to generate the dataset of reactions, which has been shown to be successful in cases when many reaction pathways have to be evaluated.[19–21] More specifically, we rely on the growing string method[22] to automatically optimize reaction paths and TSs.

We report quantum chemical data on more than 16 000 reactions in the form of reactants, products, and TSs at the B97-D3/def2-mSVP level of theory, and 12 000 reactions at the ωB97X-D3/def2-TZVP level of theory. The data include the raw output from geometry optimizations and frequency calculations in addition to atom-mapped SMILES, activation energies, and enthalpies of reaction. All reactions are gas-phase calculations involving up to seven carbon, oxygen, or nitrogen atoms per molecule. The reactants are sampled from GDB-7, a subset of GDB-17,[23] meaning that all reactions have a unimolecular reactant but potentially multi-molecular products. Figure 4.1 illustrates the dataset generation process and the resulting space of reactions in terms of their activation energies and enthalpies of reaction.

4.2 Methods

4.2.1 Overview

The dataset generation procedure started by selecting molecules from GDB-7,[23] generating conformers, and optimizing the lowest-energy conformer. An exhaustive set of driving coordinates subject to valence and connectivity constraints were generated for each reaction. Reaction paths were calculated with the growing string method,[22] which searched along each of the driving coordinates. Products and TSs discovered in this way were reoptimized, duplicate reactions were removed, and checks were performed to verify the reactions. The generated reactions were then refined at a higher level of theory. Because of the large number of density functional theory (DFT) calculations required, the massively parallel nature of the calculations was exploited by running thousands of calculations in parallel on a supercomputer.

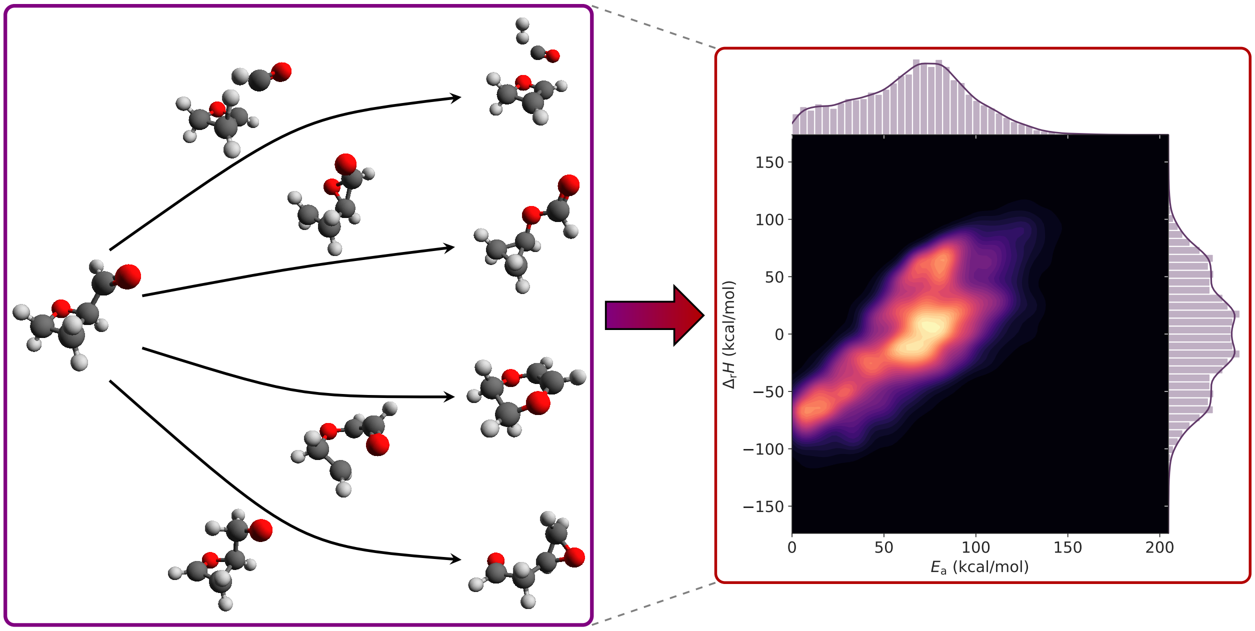


Figure 4.1. Reaction data generation and visualization of reaction space. During data generation, many reactants are optimized, hundreds of reaction paths for each reactant are searched with an automated transition state finding method, and the resulting products are optimized. The reaction space spans a wide range of activation energies and is visualized with a bivariate kernel density estimate (using a Gaussian kernel) of the probability density of the activation energy and enthalpy of reaction. The visualization encompasses both forward and reverse reactions.

4.2.2 Reactant optimization

Because of the unfavorable scaling of quantum chemical calculations, we only considered molecules with at most seven heavy atoms (C, N, O). *All* molecules with six or fewer heavy atoms were selected from GDB-7 (∼770) and a random selection of ∼430 molecules were selected from the set with seven heavy atoms. Starting from the SMILES strings, we embedded several hundred conformers for each molecule using RDKit[24] with the ETKDG distance geometry method[25] and relaxed their geometries using the MMFF94 force field implemented in RDKit. The lowest energy structure was selected for each molecule and optimized at both the B97-D3/def2-mSVP with Becke-Johnson damping level of theory[26] and the ωB97X-D3/def2-TZVP[27] level of theory with Q-Chem 5.1.[28] We ascertained that none of the molecules contained imaginary frequencies. All calculations, including the subsequent string method calculations, were done in the singlet state and used a spin-unrestricted ansatz because the bond distortions occurring in the corresponding TSs might be better treated with an unrestricted formulation. The def2-mSVP basis set in the Karlsruhe *def2* basis set family[29] is a modified version of def2-SV(P), which corrects for an overestimation of bond lengths involving hydrogen.[30] All DFT calculations used the *SG-2* standard quadrature grid, which is of sufficient quality for B97-based functionals.[31]

4.2.3 Potential energy surface exploration

The most demanding and most time-intensive step of the reaction generation process is the optimization of reaction paths to the minimum energy paths (MEPs) containing the correct TS structures. We accomplished this in an automated fashion by using the single-ended growing string method (GSM)[22] at the B97-D3/def2-mSVP level of theory. GSM performs the reaction path optimization using a set of delocalized internal coordinates, which means that the resulting MEPs may be slightly different than those obtained via a reaction path following procedure in mass-weighted internal coordinates.[32] Single-ended methods only require a reactant structure to find reactions whereas double-ended methods additionally require knowledge of the product.[33,34] *A priori* specification of the product can be problematic when there is no simple elementary step connecting reactant and product. Single-ended GSM solves this issue by only requiring a set of driving coordinates to initiate the reaction path search.

In our case, the driving coordinates are specified as bond transformations in terms of primitive internal coordinates. The direction given by the primitive internal coordinate vector is projected onto the nonredundant delocalized internal coordinates,[35] which is the space in which the reaction path optimization occurs. This results in a single tangent vector that represents all of the driving coordinates simultaneously. Importantly, this allows all other coordinates to change without constraint during the optimization, thus allowing necessary angle, torsion, and even additional bond changes to occur. Once a path has been grown, the entire path is optimized towards the MEP while monitoring the number of TSs along the path and truncating it if more than one TS is detected—ensuring that the reaction is elementary. As a result of this, not all bond changes given in the driving coordinates are guaranteed to occur. Towards the end of the path optimization, an exact TS search takes place guided by curvature information from the string.

In order to obtain many reactions, we generated an exhaustive list of driving coordinate sets for each reactant subject to a few constraints. Because elementary reactions usually involve few bond changes, we specified that at most two bonds could be broken, at most two bonds could be formed, and a total of at most three bonds could be changed. A "bond" in this sense ignored bond orders and only considered whether two atoms were connected to each other. Note that these constraints were only selected to ensure a computationally tractable number of driving coordinates. As described in the previous paragraph, these limits did not apply during the actual path optimization, they were only used to specify the initial search direction. We also ignored driving coordinates involving only a single bond change as these would likely correspond to barrierless associations or dissociations. Driving coordinates involving equivalent hydrogens were not included. Equivalent hydrogens only differ in their atom indices, e.g., a hydrogen atom that is part of a methyl group is considered to be equivalent to another hydrogen in the same methyl group. Lastly, the driving coordinates were further limited based on the valences of the expected product structures. Hydrogen atoms must have one bond, carbons can be connected to a minimum of two and a maximum of four atoms, oxygen to a minimum of one and a maximum of two, and nitrogen to a minimum of one and a maximum of three.

This process usually resulted in several hundred sets of driving coordinates per reactant. Following each GSM calculation, the endpoint of the paths were subjected to additional geometry optimizations to ensure that the product structures were at a minimum. For each reactant, there were many duplicate reactions. Instead of discarding all of them, up to four duplicates of the same reaction were retained for additional TS optimization in case some of the optimizations fail. While GSM already produces a mostly optimized TS structure, the additional optimization step ensured that the TSs were optimized to high accuracy.

4.2.4 Reaction verification and extraction

After the additional TS optimizations, duplicate reactions were filtered out again. If duplicates were present, the lowest-barrier reaction was retained. Differences in barrier height may arise due to different TS conformers. Although GSM provided an optimized MEP for each reaction, it is possible that some reactions containing incorrect transition states remained. These were filtered out according to a normal mode analysis described in the Technical Validation section.

To convert from three-dimensional geometries to SMILES,[36] connections and bond orders could be perceived with Open Babel.[37] However, there were cases where the derived bond orders were chemically unreasonable, for example, the resulting SMILES often contained adjacent radical atoms which most likely correspond to double bonds. To eliminate unreasonable structures, we converted the Open Babel molecule to InChI,[38] which only treats bond orders implicitly and resolves the issue. A downside to using InChI is that tautomers are assigned the same string, but this can be circumvented by converting to a nonstandard InChI containing a fixed-hydrogen layer. Additionally, atom ordering was lost in the InChI conversion. We reconstructed the atom map by converting to an RDKit molecule and determining the graph isomorphism between the original molecule and the RDKit molecule without considering bond orders. In the future, an alternative procedure for perceiving SMILES could be implemented based on natural bond orbital analysis.[39]

The activation energies were extracted by adding the zero-point energies from a harmonic vibrational analysis to reactant, product, and TS energies and computing the difference between resulting TS and reactant energies. Similarly, enthalpies of formation were determined based on the difference of product and reactant energies.

4.2.5 Refinement

B97-D3/def2-mSVP strikes a reasonable balance between cost and accuracy for potential energy surface exploration, but does not provide particularly accurate energies. Therefore, we refined the discovered pathways using ωB97X-D3/def2-TZVP. As mentioned earlier, reactants were already optimized with ωB97X-D3/def2-TZVP. Reactions were extracted as described in the preceding subsection, but some duplicates were retained to increase the probability of successful reoptimization. Only the duplicate with the smallest activation energy was retained in the end. Products and TSs were then reoptimized with ωB97X-D3/def2-TZVP and the final high-level reactions were extracted as before.

4.3 Data records

Q-Chem output files, extracted SMILES, activation energies, and enthalpies of formation are available for 16 365 B97-D3/def2-mSVP reactions and for 11 961 ωB97X-D3/def2-TZVP reactions.[40] The raw log files are stored in two compressed archive files, `b97d3.tar.gz` and `wb97xd3.tar.gz` for B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP data, respectively. Each archive contains a separate folder for each reaction labelled `rxn######`, where `######` denotes the reaction number padded with zeros. Within each folder are the three log files for a reaction, `r######.log` for the reactant, `p######.log` for the product, and `ts######.log` for the transition state. Each log file contains the output of a geometry optimization and harmonic vibrational analysis.

Atom-mapped SMILES, activation energies, and enthalpies of formation for each reaction are listed in the comma-separated values files `b97d3.csv` and `wb97xd3.csv` for the B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP levels of theory, respectively. The reactions are listed in the same order as the corresponding folders in the archive files. The columns in the comma-separated values files are explained in Table 4.1.

Table 4.1. A description of the columns in the comma-separated values files.

Column label	Description
`idx`	Reaction index
`rsmi`	Reactant SMILES
`psmi`	Product SMILES
`ea`	Activation energy (kcal mol^{-1})
`dh`	Enthalpy of reaction (kcal mol^{-1})

During the potential energy surface exploration, many duplicate reactions were encountered which were filtered out. Additionally, reactions that did not pass the tests described in the Technical Validation section were removed from the final list. Nonetheless, all of these calculations also produced optimized transition states, although the reactants and products were not verified for many of them and duplicate transition states exist. These data may still prove to be useful if only transition state structures are required or if additional calculations are done to obtain the corresponding reactants and products. Therefore, the log files for all successfully optimized transition states at both levels of theory are stored in `ts_with_dup_b97d3.tar.gz` and `ts_with_dup_wb97xd3.tar.gz`. There are 69 366 B97-D3/def2-mSVP transition states and 24 987 ωB97X-D3/def2-TZVP transition states.

4.4 Technical validation and analysis of reactions

Although the growing string method produces an optimized minimum energy path that should contain the correct TS in most cases, insufficient path discretization and reoptimization of TS geometries can lead to convergence failures or result in incorrect transition states. We performed several checks to filter out incorrect reactions. We ensured that all TSs have exactly one imaginary frequency. Reactions were also removed if the energy during the TS optimization changed by more than 3 kcal mol^{-1} relative to the highest energy on the growing string path. The most important check that we performed was to verify that the atomic displacements for the imaginary frequency matched the bond changes that occurred going from the proposed reactant to product. For each proposed reaction, we determined which bonds were changing in the reaction and ensured that the imaginary frequency normal mode displacements along those bonds were larger than the displacements along all the other bonds. This indicated that movement along the reaction coordinate mostly involved atoms undergoing significant change in the reaction. After all these changes, there is still the possibility that some of the transition states are incorrect. As a final check, we removed all of the reactions where the imaginary frequency of the transition state was less than 100 cm^{-1} in magnitude, as these typically correspond to conformational changes.

To avoid excessive computational cost, DFT methods had to be used to generate the reaction dataset. The functional chosen for the string method calculations, B97-D3, does not provide accu-

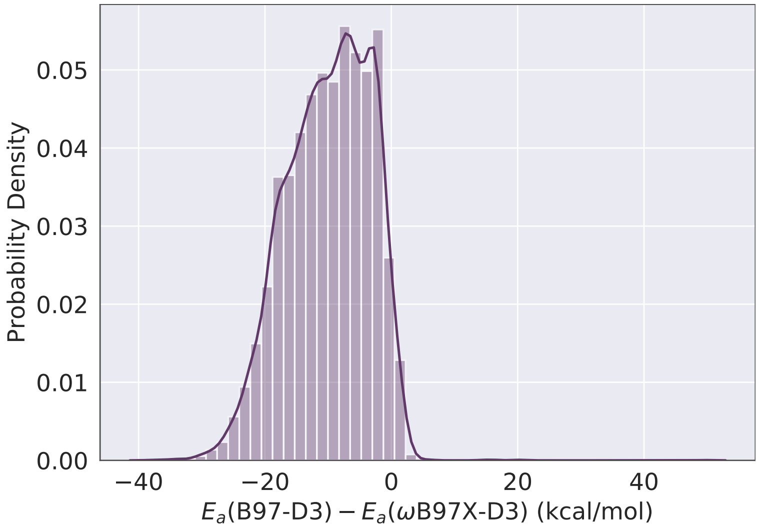


Figure 4.2. Distribution of differences between B97-D3 and ωB97X-D3 activation energies.

rate activation energies, but was selected due to its low computational cost. However, ωB97X-D3 has been shown to yield excellent quantitative barrier heights with a 2.28 kcal mol^{-1} root-mean-square deviation from reference data that is estimated to be more than ten times as accurate as the best density functionals,[41] which makes this data very useful. Figure 4.2 illustrates the striking difference between activation energies calculated with B97-D3 and those calculated with ωB97X-D3. In many cases, B97-D3 severely underestimates the barrier with an average error of 10 kcal mol^{-1}. Therefore, the following analyses were only completed for the ωB97X-D3 data.

In order to show that the dataset provides a reasonably diverse set of reactions spanning many different chemistries even though constraints were set on the number of atoms and driving coordinate generation parameters, it is necessary to characterize the types of reactions. Figure 4.1 already shows that the range of activation energies and enthalpies of formation is very large. Even high-energy reactions involving barriers of up to 200 kcal mol^{-1} are included in the dataset. If the data are used to learn reaction prediction models, including such high-energy paths is important in order to not bias models towards the low-energy regions. Figure 4.3 shows that even though the driving coordinates were limited to three bond changes, significantly more complex reactions involving more

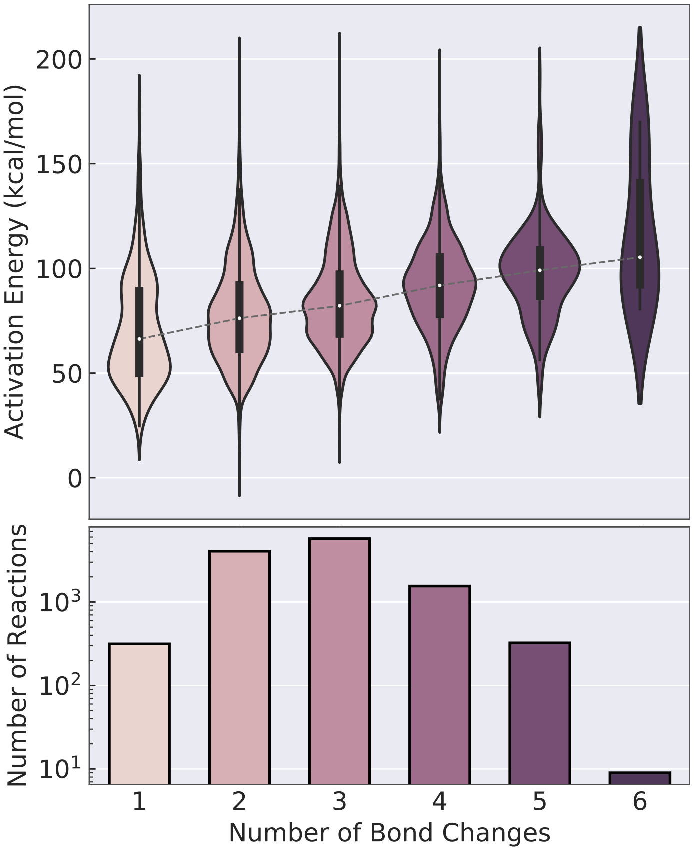


Figure 4.3. The distribution of activation energies split by the number of bond changes in the ωB97X-D3 reactions. Bond changes only consider changes in connectivity between atoms, irrespective of bond order. The distributions are scaled to have equal area.

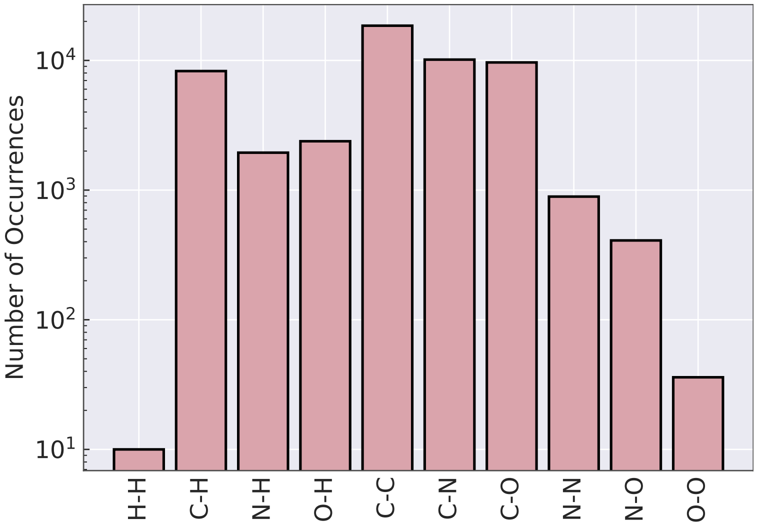


Figure 4.4. The number of times each type of bond change occurs in the ωB97X-D3 reactions. For example, C−N denotes both forming a bond between C and N atoms and breaking a bond between the atoms. This also includes a change in the bond order between the two atoms.

bond changes occur in the dataset. Nonetheless, most elementary reactions predominantly occur with only two or three bond changes. Furthermore, the median activation energy increases with an increasing number of bond changes, which is expected.

Instead of simply counting the number of bond changes, the reactions can be classified based on the types of bonds that are changed. Figure 4.4 shows that all combinations of bond changes between H, C, N, and O atoms occur in the dataset with many examples present for all reaction types. H−H changing reactions are the rarest because they only correspond to hydrogen molecule formation.

We characterized the reaction diversity by automatically extracting a set of general templates. We only focused on the reactive center by using the *GetReactingAtoms* method in RDKit to isolate atoms changing in the reaction. The molecular fragments in the reactants and products identified as the reactive center were then concatenated together to form the reaction template. In addition to the connectivity of the reacting atoms, the only features considered were atom identity, charge, aromaticity, and bond type. Figure 4.5 shows the results of this automated extraction. Many templates only have a single reaction example and only the eight most popular templates have more than 100 reaction examples, highlighting the diversity present in the dataset.

Lastly, Figure 4.6 characterizes how the pairwise atom distances change for all reactions, measured as the change going from reactant to transition state. The distribution of distance changes show that the changes for most reactions fall within a normal range, but there exists a significant tail containing some reactions with large changes. Interestingly, the second plot in Figure 4.6 shows that these large changes do not necessarily correspond to reactions with very high activation energies. However, an average increase in activation energy can be observed for those reactions with

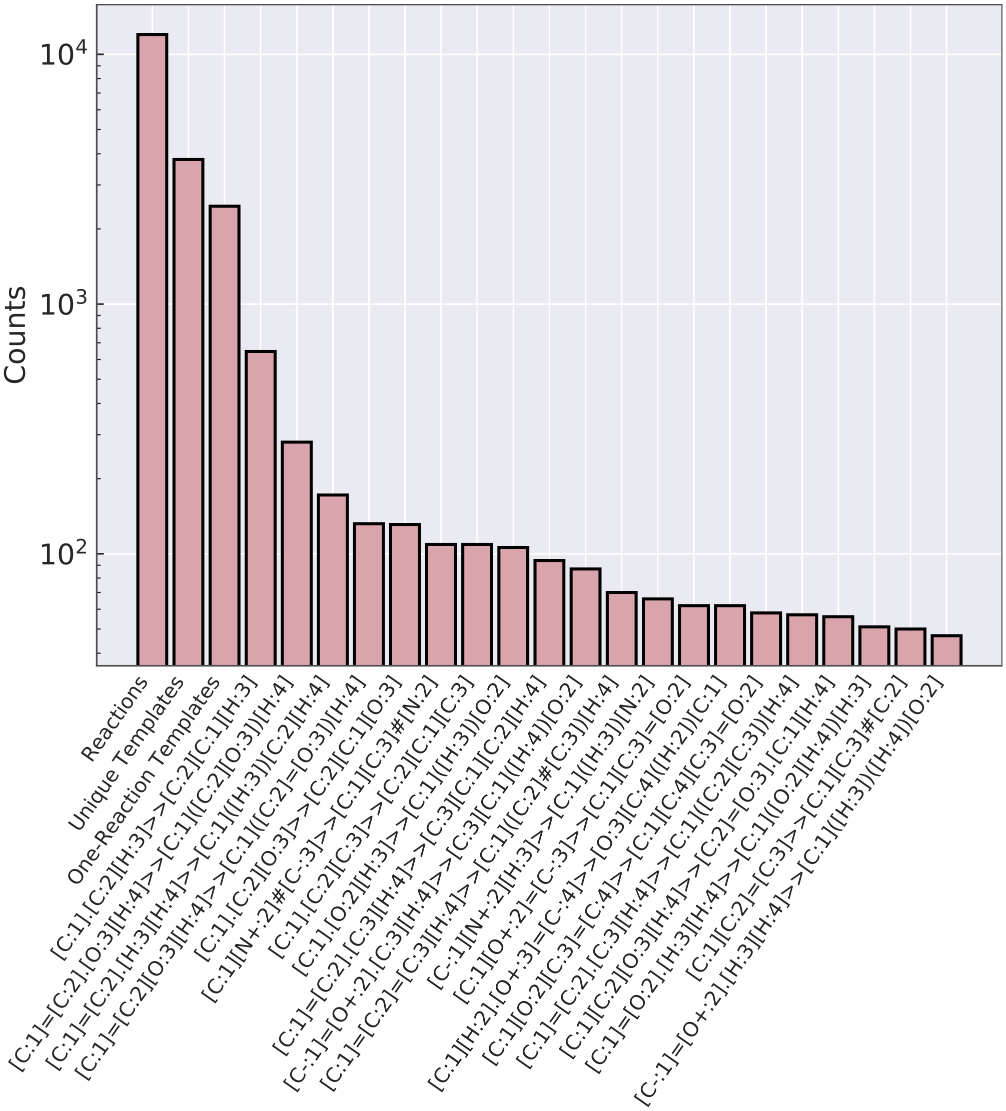


Figure 4.5. Automatically extracted reaction templates. The reactions were grouped with very general reaction templates that only consider connectivity of atoms in the reactive center, atom identity, charge, aromaticity, and bond type. The top 20 templates are denoted with SMARTS strings.[42]

normalized distance changes up to 0.3 Å.

4.5 Usage notes

With the exception of the growing string method code, which is available from the developers of the method,[43] and the Q-Chem quantum chemistry package, all code necessary to reproduce the generated data is available on GitHub.[44] The repository contains several scripts, which should be run in the following order:

- `parse_qm9.py`: Converts the QM9 data directory,[45] which contains the GDB-9 SMILES along

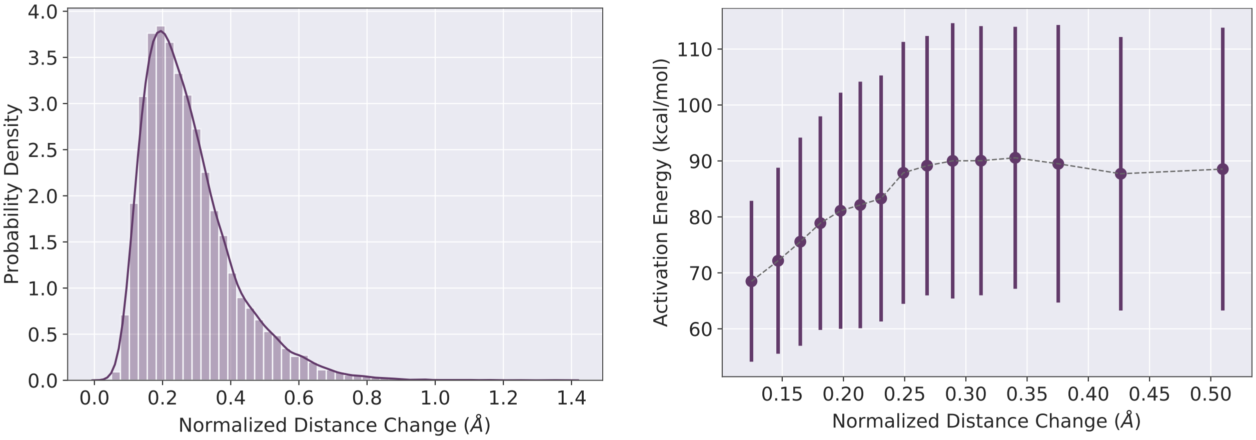


Figure 4.6. Distribution of normalized distance changes (left) and relationship to activation energy (right). The normalized distance change for each reaction is calculated by computing the pairwise distances between all atoms in the reactant and transition state. The pairwise distances are summed and normalized using the number of atoms. The distance change is then the difference between transition state and reactant. The bars in the right plot correspond to one standard deviation above and below the median of each bin.

with quantum mechanically derived properties, to a pickled file containing a list of `MolData` objects, which store the information in QM9 as Python objects.

- `make_opt_jobs.py`: Performs conformer searches and makes Q-Chem input files for optimization of reactant geometries based on the QM9 SMILES. The geometry optimizations themselves have to be performed with Q-Chem outside of the code, preferably in a massively parallel fashion on a supercomputer.

- `create_gsm_jobs.py`: Reads the geometry optimization outputs of the reactant optimizations, generates driving coordinates, and writes the files required for the GSM calculations. The GSM code has to be compiled separately.[43] The GSM calculations also have to be run separately and should produce output files with a `gsm#.out` format, where `#` corresponds to each reaction path.

- `create_prod_optfreq_jobs.py`: Reads the string endpoints from the successfully completed GSM calculations and writes the Q-Chem input files for the product optimizations.

- `create_ts_optfreq_jobs.py`: Extracts the TS geometries from the GSM output files, removes duplicate reactions using the output from the product optimizations, and writes the Q-Chem input files for additional TS optimizations.

- `extract_reactions.py`: Extracts the unique reactions using the reactant, product, and TS optimization outputs in the form of a comma-separated values file containing SMILES, activation energies, and enthalpies of reaction. Can also write the file path information of all relevant log files to the CSV output, which can be used to copy the log files for every reaction.

- `refine_reactants.py`: Writes Q-Chem input files for reoptimization of the reactants at the higher level of theory.
- `refine_products_and_ts.py`: Uses the same method as implemented in `extract_reactions.py` to extract reactions and write Q-Chem input files for the reoptimization of products and TSs at the higher level of theory. After running the Q-Chem jobs, `extract_reactions.py` can be run again to extract the high-level reactions.

If desired, the levels of theory and the reaction generation settings can be changed in the `config` folder.

4.6 References

(1) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe Jr., E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395.

(2) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'Min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(3) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(4) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495–6502.

(5) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(6) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(7) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(8) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **2020**, *71*, 121–142.

(9) Vereecken, L.; Aumont, B.; Barnes, I.; Bozzelli, J. W.; Goldman, M. J.; Green, W. H.; Madronich, S.; Mcgillen, M. R.; Mellouki, A.; Orlando, J. J.; Picquet-Varrault, B.; Rickard, A. R.; Stockwell, W. R.; Wallington, T. J.; Carter, W. P. Perspective on Mechanism Development and Structure-Activity Relationships for Gas-Phase Atmospheric Chemistry. *Int. J. Chem. Kinet.* **2018**, *50*, 435–469.

(10) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(11) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(12) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(13) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(14) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C−N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186–190.

(15) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004–5008.

(16) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D., The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information In *The Future of the History of Chemical Information*; American Chemical Society: Washington, DC, 2014; Chapter 8, pp 127–148.

(17) Mayfield, J.; Lowe, D.; Sayle, R., Pistachio: Search and Faceting of Large Reaction Databases In *American Chemical Society National Meeting*, Washington, DC, 2017.

(18) Lowe, D. Chemical Reactions from US Patents (1976-Sep2016), Figshare. https://doi.org/10.6084/m9.figshare.5104873.v1, 2017.

(19) Zádor, J.; Miller, J. A. Adventures on the C_3H_5O Potential Energy Surface: OH + Propyne, OH + Allene and Related Reactions. *Proc. Combust. Inst.* **2015**, *35*, 181–188.

(20) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *8*, e1354.

(21) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zádor, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(22) Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

(23) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(24) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed 10/06/2019).

(25) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(26) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(27) Lin, Y. S.; Li, G. D.; Mao, S. P.; Chai, J. D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(28) Shao, Y. et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(29) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Balence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(30) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent Structures and Interactions by Density Functional Theory with Small Atomic Orbital Basis Sets. *J. Chem. Phys.* **2015**, *143*, 054107.

(31) Dasgupta, S.; Herbert, J. M. Standard Grids for High-Precision Integration of Modern Density Functionals: SG-2 and SG-3. *J. Comput. Chem.* **2017**, *38*, 869–882.

(32) Gonzalez, C.; Schlegel, H. B. Reaction Path Following in Mass-Weighted Internal Coordinates. *J. Phys. Chem.* **1990**, *94*, 5523–5527.

(33) Zimmerman, P. Reliable Transition State Searches Integrated with the Growing String Method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(34) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(35) Baker, J.; Kessi, A.; Delley, B. The Generation and Use of Delocalized Internal Coordinates in Geometry Optimization. *J. Chem. Phys.* **1996**, *105*, 192–212.

(36) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(37) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*.

(38) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*.

(39) Weinhold, F.; Landis, C. R.; Glendening, E. D. What is NBO Analysis and How Is It Useful? *Int. Rev. Phys. Chem.* **2016**, *35*, 399–440.

(40) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry, version 1.0.1, Zenodo. https://doi.org/10.5281/zenodo.3581266, 2020.

(41) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(42) Daylight Chemical Information Systems, Inc. SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 11/21/2019).

(43) Zimmerman, P. molecularGSM. https://github.com/ZimmermanGroup/molecularGSM (accessed 06/05/2019).

(44) Grambow, C. A. cgrambow/ard_gsm: Release version 1.0.0, version 1.0.0, Zenodo. https://doi.org/10.5281/zenodo.3552859, 2019.

(45) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

Chapter 5

Deep learning of activation energies

Much of this work has previously appeared as Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997. Lagnajit Pattanaik assisted with extracting the reaction types by providing code to characterize the bond changes of each reaction. The code for training and evaluating the machine learning models is available at `https://github.com/cgrambow/chemprop/tree/reaction`.

5.1 Introduction

Activation energy is an important kinetic parameter that enables quantitative ranking of reactions for automated reaction mechanism generation and organic synthesis planning. Achieving reliable activation energy prediction is an integral step toward the complete prediction of kinetics. Machine learning, particularly deep learning, has recently emerged as a promising data-driven approach for reaction outcome prediction[1–6] and for use in organic retrosynthetic analysis.[7–10] These methods leverage massive data sets of organic reactions, such as Reaxys[11] and Pistachio.[12] However, the methods operate on qualitative data that indicate only the major reaction product and mostly lack any information regarding reaction rates. Moreover, many of the organic chemistry reactions are not elementary. The data used by Kayala and Baldi[1] and Fooshee et al.[2] are an exception, but quantitative information is still missing.

Linear methods, like Evans-Polanyi relationships[13] and group additivity models,[14–17] have been successfully used in automated reaction mechanism generation to estimate rate constants, but they are limited in scope and applicability. New parameters have to be derived for each reaction family, and predictions far from the training data often go awry. Nonlinear decision trees provide more flexible models for the estimation of kinetics but are also most effective when they are specific to a reaction family.[18] Neural networks may be a promising alternative as large data sets become more readily available.

Recently, some quantitative reaction prediction research using neural networks has become available, but it is limited in its application. Gastegger and Marquetand developed a neural network potential for a specific organic reaction involving bond breaking and formation, likely the first of its kind.[19] Allison described a rate constant predictor for a specific reaction type involving reactions with OH radicals.[20] Choi et al. looked specifically at activation energy prediction using machine learning.[21] However, their training data were composed of reactions in the Reaction Mechanism Generator (RMG)[18] database that comprised many similar reactions such that a random test split yielded ostensibly good results. Their issue stems from the fact that the vast majority of the RMG database is composed of just two reaction families: hydrogen abstraction and addition of a radical to a multiple bond. Reactions within the same family tend to have similar kinetics. Therefore, a model trained on such data performs particularly well for the two reaction types but not well for others. Moreover, the model of Choi et al. required knowledge of the reaction enthalpy and entropy to make a prediction. Singh et al. similarly predicted reaction barriers for a small data set of dehydrogenation reactions involving dissociation of N_2 and O_2 on surfaces.[22] Their model also required the reaction energy as additional input.

Our goal is to develop a deep learning model to predict activation energies across a wide range of reaction types that does not depend on any additional input and requires only a graph representation of reactants and products. Such a model would be useful as a first step in deep learning-based estimation of kinetics for automated reaction mechanism generation (e.g., in RMG[18]) or would allow for quantitative ranking of reaction candidates that were generated via combinatorial enumeration of potential products given a reactant.[23] Training such a model requires suitable quantitative data. We use data based on large-scale quantum chemistry calculations,[24] but high-throughput experimentation[25] is also starting to become a valuable source of new data.

5.2 Methods

To effectively learn activation energy, we must encode the atoms involved in a reaction that change significantly in a way that they contribute most to the predicted property. To accomplish this, we extend a state-of-the-art molecular property prediction method, `chemprop`, developed by Yang et al.,[26] to work with atom-mapped reactions. Figure 5.1 shows a schematic of the method, which will be explained in the following. Our modified code is available on GitHub in the `reaction` branch of `chemprop`.[27] The code also includes the trained model in the `model` directory that can be directly used with the `predict.py` script in `chemprop`.

5.2.1 Neural network architecture

The method of Yang et al. is a directed message passing neural network (D-MPNN) for molecular property prediction, which is a type of graph convolutional neural network.[28–30] Graphs naturally represent molecules in which atoms are the vertices and bonds are the edges. Our method extends

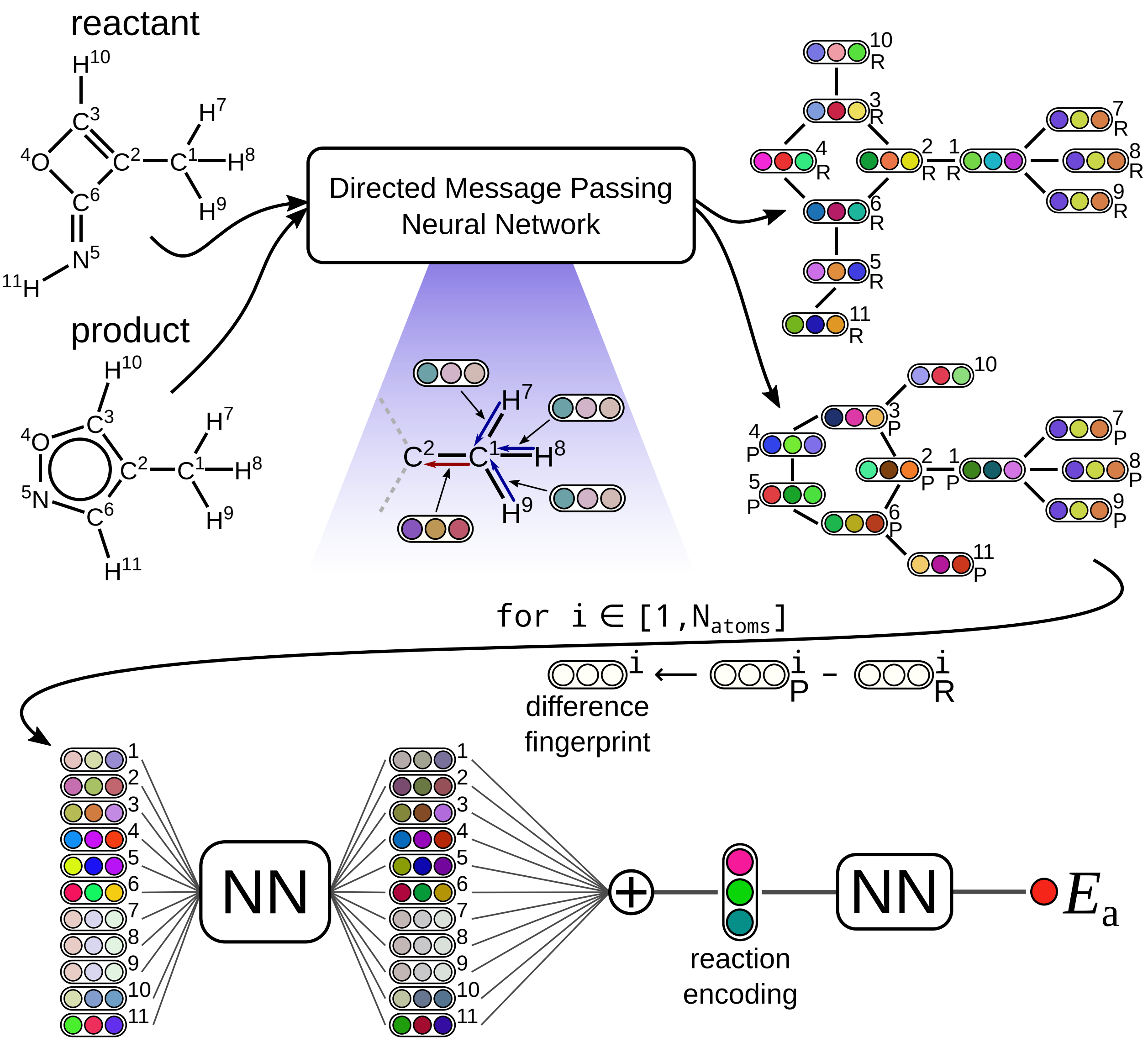


Figure 5.1. Illustration of neural network model used to predict activation energy given a chemical reaction. The difference fingerprint approach emphasizes atoms that change during the reaction.

the `chemprop` architecture in order to encode reactions instead of molecules by using a difference fingerprint approach, which is similar to that of Coley et al..[5]

The same D-MPNN operates on the reactant and product graphs, G_R and G_P, separately to create a learned representation for each atom in the reactant and each atom in the product. Hydrogens are explicit in the graphs because they are often directly involved in the reactions. We subtract the representations for corresponding atoms between reactants and products from each other to generate a reaction embedding for each atom. We then aggregate these embeddings prior to the final activation energy prediction.

To operate on a molecular graph, $G = (V, E)$ with vertices (atoms) V and edges (bonds) E, we require initial atom features $\{x_v \mid v \in V\}$ and initial bond features $\{e_{vw} \mid vw \in E\}$. The atom features comprise a one-hot encoding of the atomic number, the degree, the formal charge, the chiral tag, the total number of hydrogens, and the hybridization; an aromaticity flag; the atomic mass; and whether the atom is in a ring of size s for $s \in [3, 10]$. The bond features indicate whether the bond is a single, double, triple, or aromatic bond; whether it is conjugated; whether it is in a ring; whether it is in a ring of size s for $s \in [3, 10]$; and they contain a one-hot encoding of the bond stereochemistry. Since the ring membership features for atoms and bonds are one-hot vectors, they are able to encode all different-size rings that they are part of. We obtained all of the features using RDKit.[31]

The following illustrates the message passing procedure. Note that some layers may include bias parameters, but the equations do not show them explicitly. We obtain the initial hidden state of a bond vw in an embedding operation given by

$$h^0_{vw} = \tau \left(W_i \ \mathtt{cat}(x_v, e_{vw}) \right) \tag{5.1}$$

where $\tau(\cdot)$ is the ReLU activation function, $W_i \in \mathbb{R}^{h \times (h_x + h_e)}$ is a learned matrix, and $\mathtt{cat}(x_v, e_{vw}) \in \mathbb{R}^{h_x + h_e}$ represents the concatenation of atom and bond features. h_x and h_e are the sizes of the initial atom and bond features, respectively. We determined the optimal hidden size to be $h = 1800$ using the hyperparameter optimization procedure described later. The network calculates messages at the next time step as

$$m^{t+1}_{vw} = \sum_{k \in \{N(v) \setminus w\}} h^t_{kv} \tag{5.2}$$

where $N(v)$ denotes the neighbors of atom v. The hidden state is updated by

$$h^{t+1}_{vw} = \tau \left(h^0_{vw} + W_m m^{t+1}_{vw} \right) \tag{5.3}$$

where $W_m \in \mathbb{R}^{h \times h}$ is another learned matrix and adding h^0_{vw} connects every hidden state to its original embedding. This proceeds iteratively for $t \in \{1, \ldots, T\}$, and we set $T = 5$. We then convert bond fingerprints to atom fingerprints according to

$$m_v = \sum_{w \in N(v)} h^T_{wv} \tag{5.4}$$

$$h_v = \tau \left(W_a \ \mathtt{cat}(x_v, m_v) \right) \tag{5.5}$$

where $W_a \in \mathbb{R}^{h \times (h_x + h)}$ is a third learned matrix. Equations (5.4) and (5.5) are another message passing step, so the total number of message passing iterations is $T + 1 = 6$. We apply the operations in Equations (5.1) to (5.5) to both the reactant and the product to yield $h^{(R)}_v$ and $h^{(P)}_v$, respectively, for all atoms v in the molecular graph.

Next, we obtain the embedded difference atom fingerprints as

$$d_v = \tau \left(W_d \left(h_v^{(P)} - h_v^{(R)} \right) \right) \tag{5.6}$$

where $W_d \in \mathbb{R}^{h \times h}$ is a learned matrix. We sum the difference fingerprints to obtain a feature vector for the reaction

$$r = \sum_{v \in G} d_v \tag{5.7}$$

Before generating an estimate for the activation energy, we calculate 200 global molecular features using RDKit[31] for both the product and the reactant and append their difference to the reaction feature vector

$$\tilde{r} = \texttt{cat}(r, f_P - f_R) \tag{5.8}$$

where f_P and f_R are the product and reactant RDKit features, respectively. The purpose of these features is to capture global structural information in addition to the local information that is built up in the message passing steps. See Ref. [26] for more information.

Finally, the reaction feature vector with a linear activation enables estimation of the activation energy

$$\widehat{E}_a = w_a^\intercal \tilde{r} \tag{5.9}$$

where $w_a \in \mathbb{R}^{h+200}$ is a learned vector. We observed that a multitask prediction of both the activation energy and the enthalpy of reaction significantly improves the activation energy estimate. Therefore, the model has a second output to predict the enthalpy of reaction

$$\Delta \widehat{H}_r = w_e^\intercal \tilde{r} \tag{5.10}$$

which is supplied during training but no longer used during evaluation.

The idea behind constructing difference fingerprints is to subtract out the effects of atoms that do not change significantly in the reaction, and, therefore, do not contribute much to the activation energy prediction. This requires that atom-mapped reactions are available, which is often not the case, but developing methods for automatic atom mapping is an active area of research.[32–34] With large molecules, even atoms that do not change their covalent bonding environment may have large difference fingerprint magnitudes because they may strengthen van der Waals attractions between different parts of the molecule or sterically hinder certain transition states.

5.2.2 Data sources

We train our model on a newly developed gas-phase organic chemistry data set of elementary atom-mapped reactions based on density functional theory (DFT).[24] These data span a diverse set of reactions with at most seven heavy atoms involving carbon, hydrogen, nitrogen, and oxygen. Reactions are available at two different levels of theory: B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP. Both levels are used in a transfer learning approach similar to that in Ref. [35], but we measure the final model performance against the ωB97X-D3/def2-TZVP data. We augment our data by including all of the reverse reactions, as well, which essentially doubles the training data and may further help in subtracting out the effect of distant atoms. This results in a total of 33 000 B97-D3/def2-mSVP and 24 000 ωB97X-D3/def2-TZVP reactions. The activation energies, E_a, provided in the data set are *not* obtained by fitting to an Arrhenius form, but they represent the difference of transition state and reactant electronic energies, each including zero-point energy.

The reactions in automated reaction mechanism generation often involve radicals because large chemical mechanisms are usually driven by radical propagation reactions. We use the same methodology as described in Chapter 4, using an unrestricted DFT formalism, to compute an additional 2024 B97-D3/def2-mSVP and 1367 ωB97X-D3/def2-TZVP reactions[24] starting from a random selection of reactants using the data published by St. John et al..[36] As described in Chapter 4, we ensured that the normal mode displacements for the mode corresponding the single imaginary frequency of each transition state are congruent with the bond changes of the reaction. The set of radical reactions is available online.[37] As with the non-radical reactions, the reverse reactions are used to augment the radical reaction data.

5.2.3 Training and hyperparameter optimization

We partition the data into training, validation, and testing sets using a scaffold split, which bins the data based on the Murcko scaffolds of the reactants calculated by RDKit.[31] Ref. [26] describes the exact partitioning procedure. To obtain a better measure of model performance, we use a 10-fold cross-validation approach. The validation data sets, used for hyperparameter optimization and early stopping, consist of 5% of the available data. Even though the model produces $\widehat{E}_a$ and $\Delta\widehat{H}_r$ as outputs, we only use the error in $\widehat{E}_a$ to determine early stopping. We schedule the learning rate as follows: a linear learning rate increase from the initial learning rate to the maximum learning rate over a given number of warm-up epochs followed by an exponential decrease to the final learning rate over the course of the remaining epochs.

Training proceeds in two parts. First, we train the base model with the low-level B97-D3/def2-mSVP data. We then initialize the parameters of the final model using those of the base model and train the final model on the high-level ωB97X-D3/def2-TZVP data. This transfer learning approach makes better use of all available data and enables improved accuracy of the final model.

We determin the architecture and other hyperparameters using the hyperparameter optimization code supplied with the `chemprop` package.[26] In addition to the hidden size, h, and other architectural parameters, we optimize several training hyperparameters including the batch size, the number of

Table 5.1. Optimized training hyperparameters.

Hyperparameter	Base Model	Final Model
Batch size	50	10
Number of epochs	80	60
Initial learning rate	10^{-5}	10^{-4}
Maximum learning rate	10^{-3}	10^{-4}
Final learning rate	10^{-5}	10^{-6}
Number of warm-up epochs	3	1

epochs, the initial learning rate, the maximum learning rate, the final learning rate, and the number of warmup epochs. Table 5.1 shows the optimized parameters.

5.3 Results and discussion

To assess whether the trained model can make useful predictions across a wide range of chemical reactions, the test set should contain reactions that are sufficiently different from those in the training data, i.e., out-of-domain data. To generate such a data split, we partitioned our data on the basis of the scaffold splitting technique, which has been shown to approximate time splits that are common in industry and are a better measure of generalizability than random splits.[26] We performed the split on the basis of the scaffold of the reactant molecule. Moreover, to obtain a less variable estimate of the model performance, we evaluated the model using 10-fold cross-validation. A split into 85% training, 5% validation, and 10% test data yields a test set mean absolute error (MAE) of $1.7 \pm 0.1\,\mathrm{kcal\,mol^{-1}}$ and a root-mean-square error (RMSE) of $3.4 \pm 0.3\,\mathrm{kcal\,mol^{-1}}$, where the indicated bounds correspond to one standard deviation evaluated across the ten folds. While this error is quite small given the diverse nature of the data, which span an activation energy range of $200\,\mathrm{kcal\,mol^{-1}}$,[24] the true error is confounded by the accuracy of the ωB97X-D3 method used to generate the training and test data, which itself has an RMSE of $2.28\,\mathrm{kcal\,mol^{-1}}$ measured against much more accurate reference data.[38]

Because the model does not take three-dimensional structural information into account and because the training and test sets include only a single conformer for each molecule (not necessarily the most stable one), some of the error is due to conformational variations of the reactant or product structures. More accurate models could be based on the molecular geometries, which have been shown to work well for molecular property prediction and the development of neural network potentials.[39] Nonetheless, we do not employ such information here because it is often not readily available in applications when one wishes to rapidly predict activation energies, like in automated reaction mechanism generation.

5.3.1 Error analysis

More fine-grained results are shown in Figure 5.2. The parity plot in Figure 5.2a shows that accurate predictions are made across the entire activation energy range, and this accuracy is even maintained in regions where data are sparser. Furthermore, there seems to be no systematic over- or underprediction, and large outliers are relatively infrequent. This is further shown in the error histogram in Figure 5.2b, which indicates that only very few reactions have errors in excess of $10\,\mathrm{kcal\,mol^{-1}}$. Depending on the application, the model may be sufficiently accurate for quantitative predictions if errors slightly in excess of those of the ωB97X-D3 method are acceptable. An MAE of $1.7\,\mathrm{kcal\,mol^{-1}}$ implies that rate coefficients differ by a factor of 2.4 from the true/DFT value at 1000 K on average, which is often quite acceptable. However, this error increases to a factor of 17.5 at 300 K. Moreover, entropic effects that would typically be captured in the pre-exponential factor used in Arrhenius expressions are not taken into account in this analysis and would constitute an additional source of error.

Regardless, the results show that the model is suitable for ranking a list of candidate reactions by their likelihood of occurring. This may lead to an improvement over qualitative reaction outcome prediction approaches by enabling a more quantitative ranking. However, a direct comparison is not currently possible because such approaches are generally not based on elementary reactions and involve multiple steps in solvated environments. A promising, immediate application of the model could be to enable discovery of novel reactions from species in large chemical mechanisms. Reaction candidates can be generated from each molecule in a mechanism by changing bonds systematically to enumerate potential products.[23] The deep learning model can then assess which candidates have the lowest barriers and warrant further evaluation. Such a reaction discovery process would proceed in a template-free manner, whereas conventional reaction mechanism generation software is based on templates to restrict the allowable chemistry.[18]

Figure 5.2c also shows that the model strongly benefits from additional training data, and the typical decrease in the slope of the learning curve is not yet evident. However, this is partially because hyperparameter optimization was performed on an 85:5:10 split. Optimization for the points at lower training ratios would lead to improved performance and show a more typical curve.

Unlike our model, other methods for the estimation of activation energy and kinetics, such as the decision tree estimator used in the RMG software,[18] are often applicable only within a specific reaction family/template. The decision tree templates implemented in RMG are based on known reactivity accumulated over decades and manually curated into reaction rules. Conversely, the training data for the deep learning model are obtained from systematic potential energy surface exploration and contain many unexpected reactions that do not fall within the space encoded by the RMG templates. In fact, only 15% of all reactions in the data used for this study have a matching template in RMG (shown in Figure 5.3). There is no statistically significant difference between

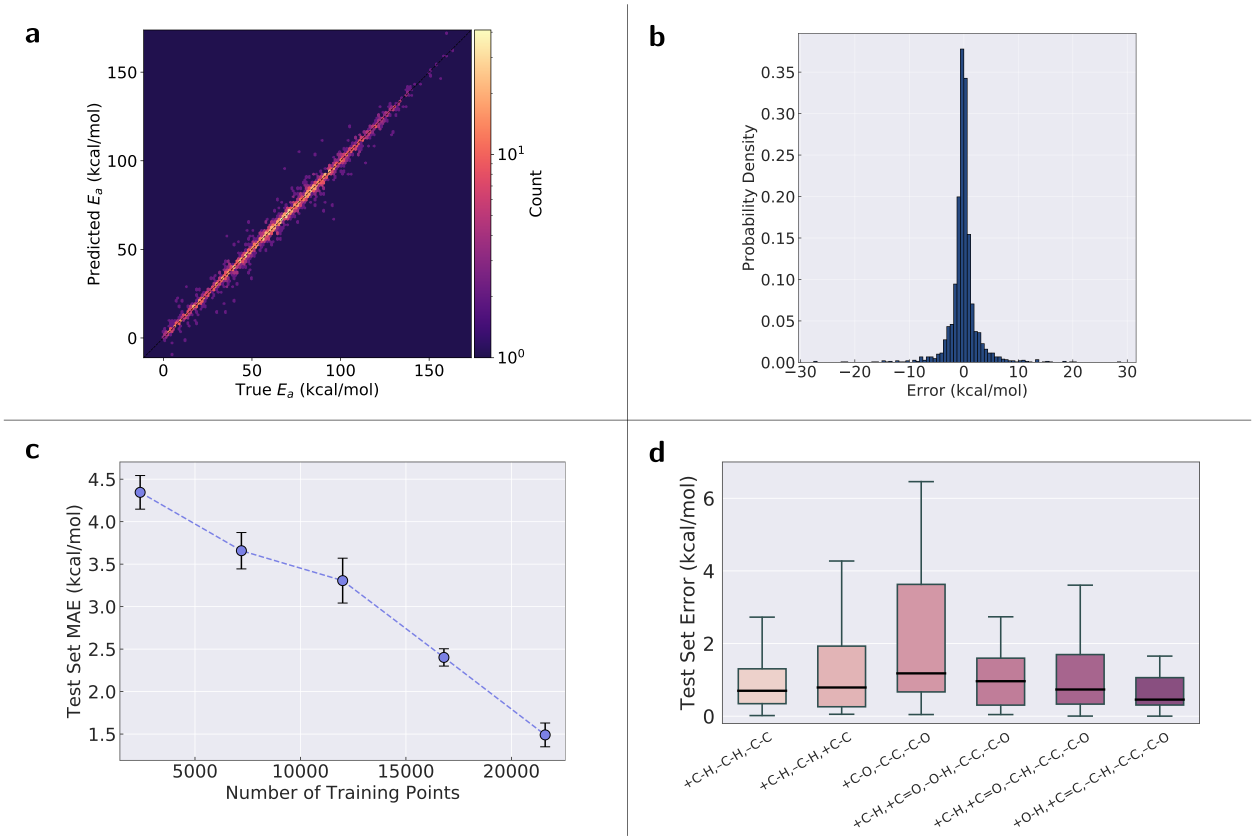


Figure 5.2. Deep learning model results. (**a**) Parity plot of model predictions vs. "true" (ωB97X-D3) data for the first fold. (**b**) Histogram of prediction errors (predicted minus "true") for the first fold. (**c**) MAE vs. the number of training data points for the deep learning model. The error bars indicate one standard deviation calculated across the ten folds. (**d**) Distributions of errors (outliers not shown) for the six most frequent reaction types (first fold). Each reaction type includes only the bond changes occurring in the reaction, e.g., +C-H,–C-H,–C-C means that a carbon-hydrogen bond is formed, a different carbon-hydrogen bond is broken, and a carbon-carbon single bond is broken in the reaction.

the deep learning model performance on RMG-type reactions and on non-RMG-type reactions ($p \leq 0.05$), which shows that our template-free model can be applied to many reactions that do not fit into expected reaction families and may be useful for discovering new and unexpected reactions.

Figure 5.2d illustrates that the test set error is not the same across all reaction types (examples of each reaction type are shown in Figures 5.4 to 5.9), but the reasons for this are not obvious. The +C-H,–C-H,–C-C type leads to the formation of carbenes via hydrogen transfer and ring opening and has a distribution of errors similar to that of the +C-H,–C-H,+C-C type, which is its reverse. Of the most frequent reaction types, the largest errors are associated with the +C-O,–C-C,–C-O type, which is similar to the +C-H,–C-H,–C-C type but involves the transfer of a hydroxy group instead of a hydrogen or the rearrangement of a cyclic ether. The last three reaction types shown in Figure 5.2d generally have small errors, although the +C-H,+C=O,–C-H,–C-C,–C-O type has a tail skewed toward larger errors, potentially because of its unusual zwitterion/carbene product. Interestingly, the model generally performs poorly for reactions with three heavy atoms as shown

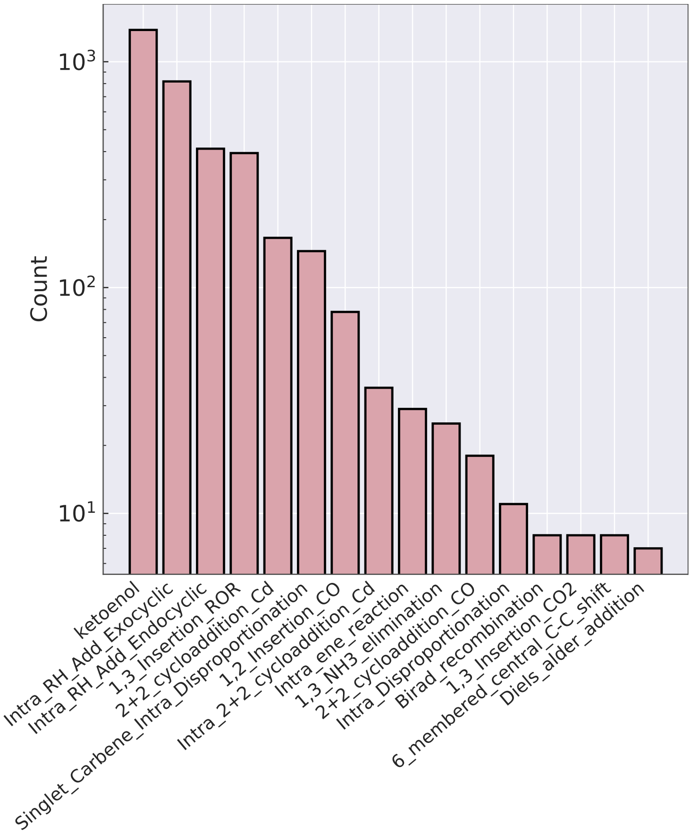


Figure 5.3. Number of reactions that match each RMG family.

in Figure 5.10, perhaps because the training data are dominated by larger molecules.

5.3.2 Reaction embeddings

To assess whether the reaction encoding learned by the model is chemically reasonable, we embedded the encodings in two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE).[40] The activation energy gradient in Figure 5.11 demonstrates that the model has learned to organize reactions it has not seen during training in a sensible manner that correlates with reactivity. Moreover, different regions in this representation of reaction space correspond to different reaction types. The six most frequent reaction types (same as those in Figure 5.2d) are highlighted in Figure 5.11. Because the reaction types are based only on the bond changes, the reactions within each type involve many different chemical functionalities; still, the model learns to cluster reactions of the same type together. The same analysis is conducted using principal component analysis (PCA) in Figure 5.12, but the separation into reaction-type clusters is not as striking because the first two PCA components capture only 46% of the total variance. Nonetheless, the clustering is still evident and a method like PCA allows new reaction samples to be transformed into the embedded space whereas t-SNE does not.

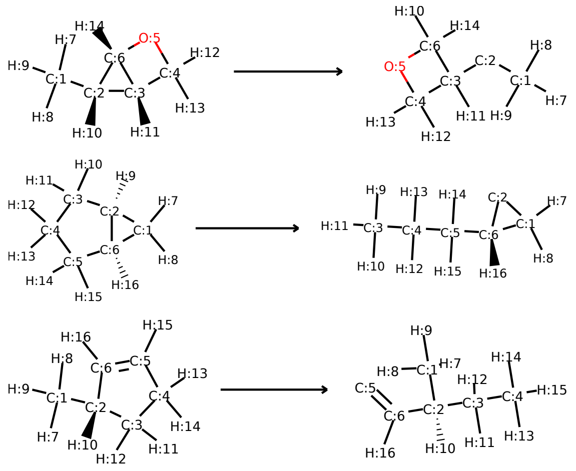


Figure 5.4. +C-H,–C-H,–C-C type reactions.

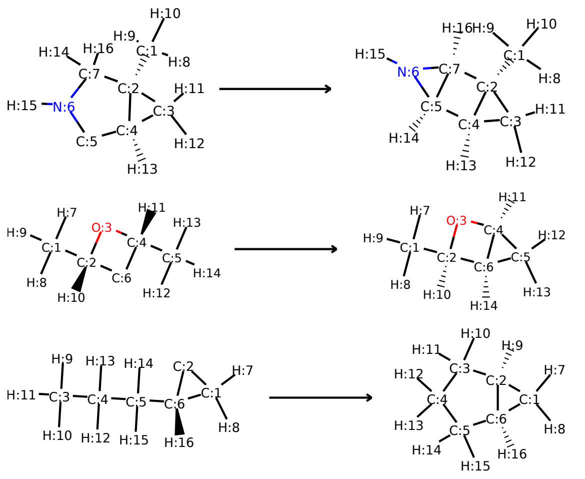


Figure 5.5. +C-H,–C-H,+C-C type reactions.

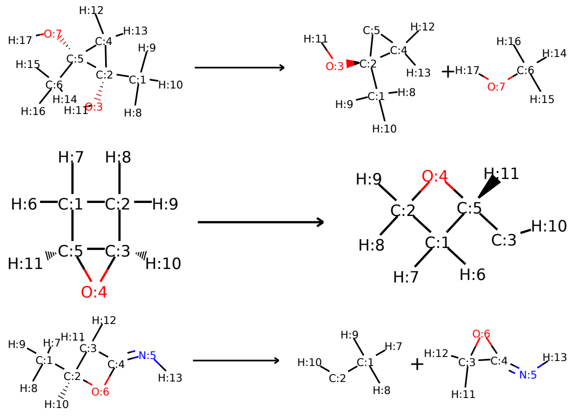


Figure 5.6. +C-O,–C-C,–C-O type reactions.

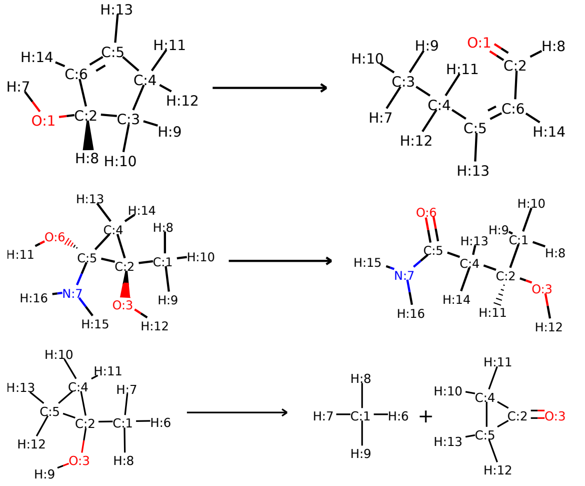


Figure 5.7. +C-H,+C=O,–O-H,–C-C,–C-O type reactions.

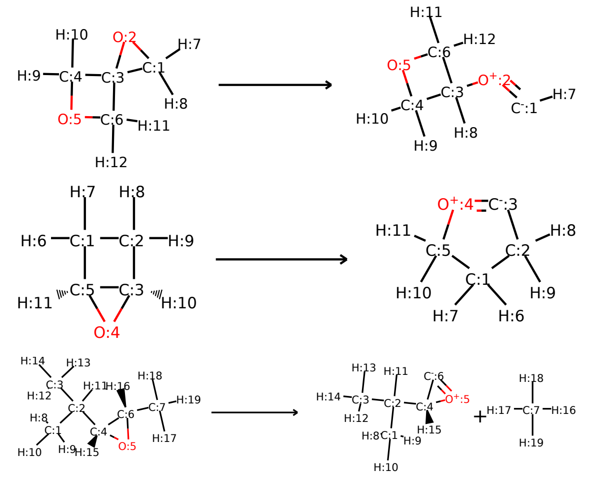


Figure 5.8. +C-H,+C=O,–C-H,–C-C,–C-O type reactions.

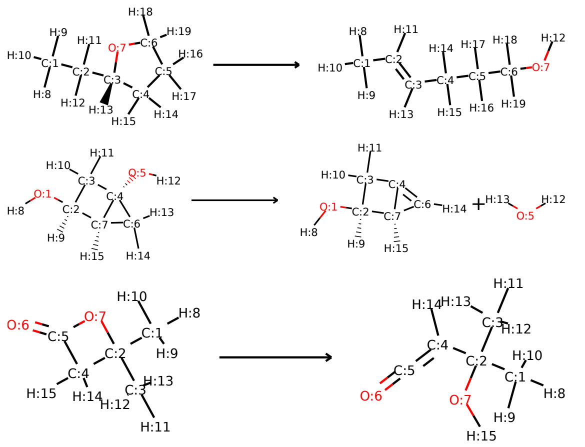


Figure 5.9. +O-H,+C=C,–C-H,–C-C,–C-O type reactions.

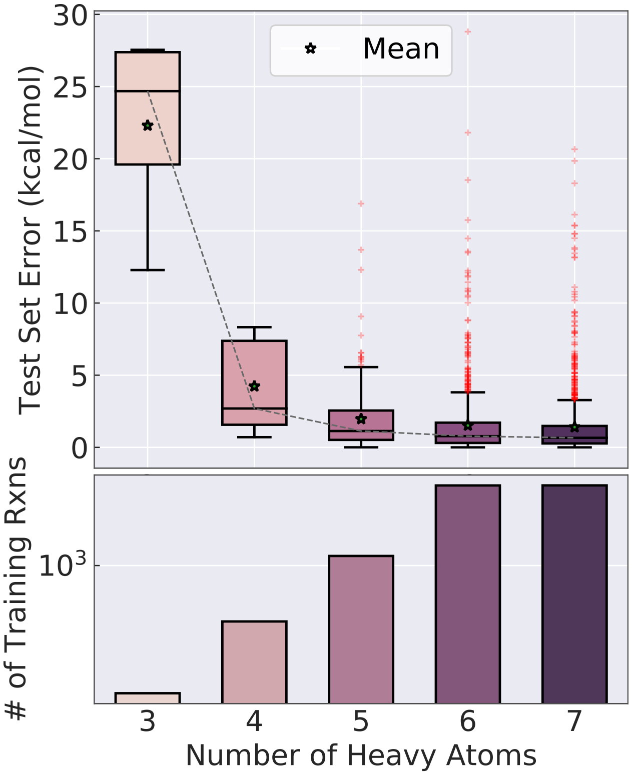


Figure 5.10. MAE split by the number of heavy atoms involved in each reaction.

5.3.3 Side chain analysis

To further show that the model behaves correctly when different functional groups are present, we analyzed the effects of substituting hydrogen atoms with side chains containing different functional groups and verified the model predictions using DFT. The analysis was conducted by selecting two reactions, one with a substitutable hydrogen close to the reaction center (at a distance of 1) and one with a substitutable hydrogen far from the reaction center (at a distance of 3), and substituting the hydrogens using different functional groups (side chains). The groups were chosen as the homologous methyl, ethyl, and propyl chains; an amino group; and a hydroxy group. Figure 5.13 illustrates the original and substituted reactions.

As shown in Figure 5.14, both the amino and hydroxy groups have a significant negative effect on the activation energy when the substitution occurs close to the reactive center. Interestingly, the more electronegative hydroxy group does not reduce the barrier as strongly as the amino group. The deep learning model agrees well with the DFT calculations, except in the case of the hydroxy group, where it predicts a barrier lower than that for the amino group. When the substitution occurs far from the reactive center, none of the side chains results in significant differences from the original barrier; and the deep learning predictions agree well with the DFT results.

This side chain analysis also agrees with the earlier hypothesis that the difference fingerprints (recall Figure 5.1) should, on average, have a smaller contribution to the activation energy for

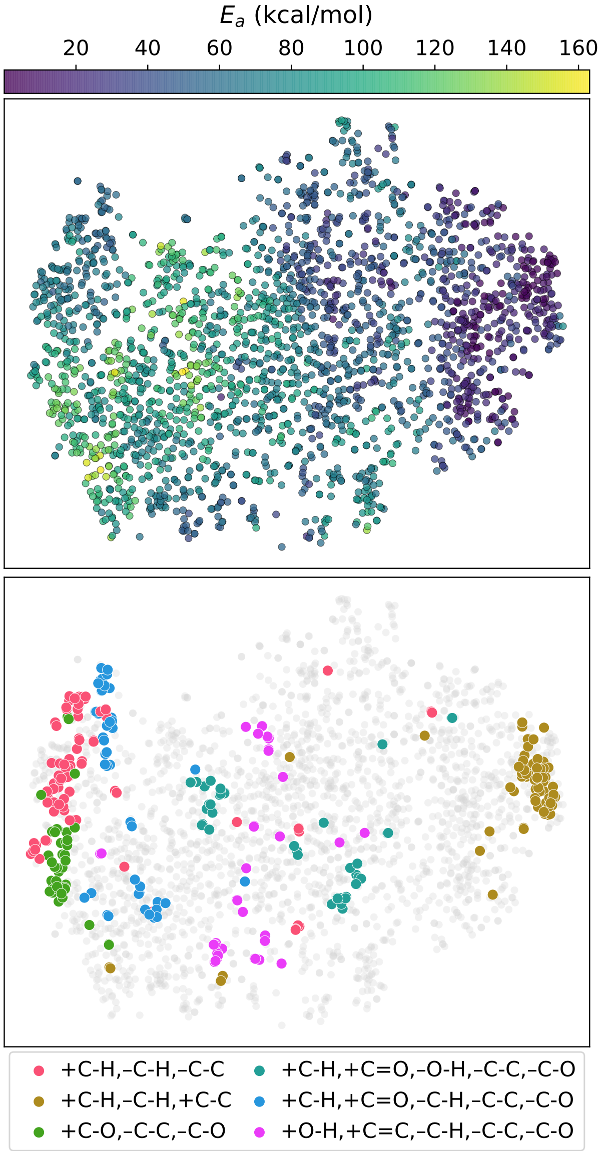


Figure 5.11. t-Distributed stochastic neighbor embedding (t-SNE) of the learned reaction encodings for the test set of the first fold. The embeddings correlate with the activation energy (top). The reactions cluster in t-SNE space on the basis of their reaction type. Shown are the six most frequent reaction types (bottom). Each reaction type includes only the bond changes occurring in the reaction, e.g., +C-H,–C-H,–C-C means that a carbon-hydrogen bond is formed, a different carbon-hydrogen bond is broken, and a carbon-carbon single bond is broken in the reaction.

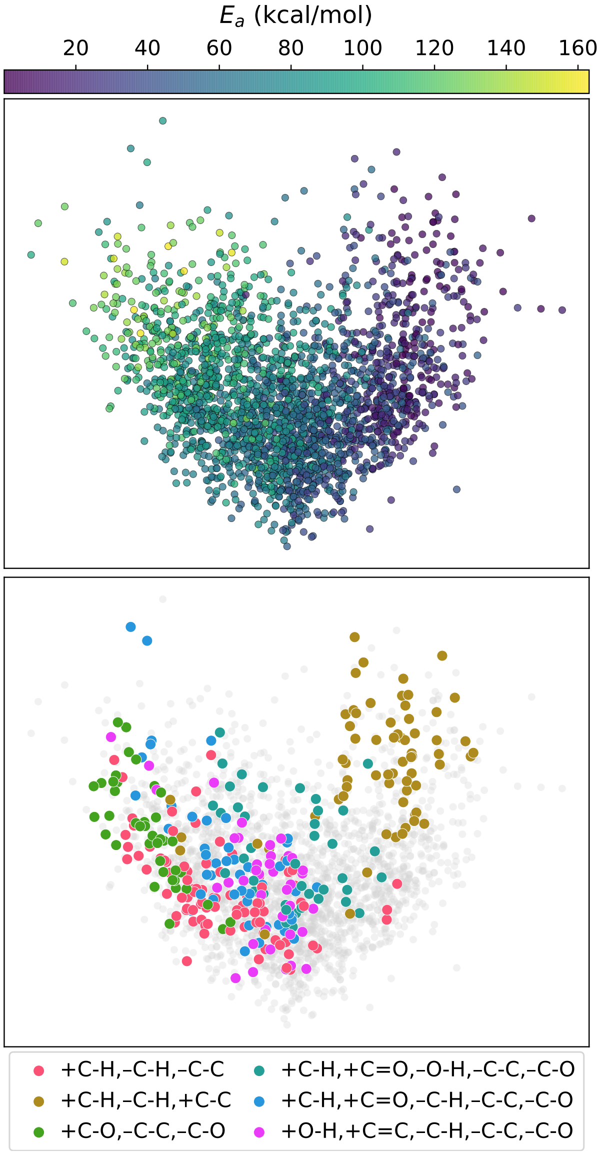


Figure 5.12. Principal component analysis (PCA) of the learned reaction encodings for the test set of the first fold. The first two components capture 46% of the total variance. The reactions cluster in PCA space on the basis of their reaction type. Shown are the six most frequent reaction types (bottom). Each reaction type includes only the bond changes occurring in the reaction, e.g., +C-H,–C-H,–C-C means that a carbon-hydrogen bond is formed, a different carbon-hydrogen bond is broken, and a carbon-carbon single bond is formed in the reaction.

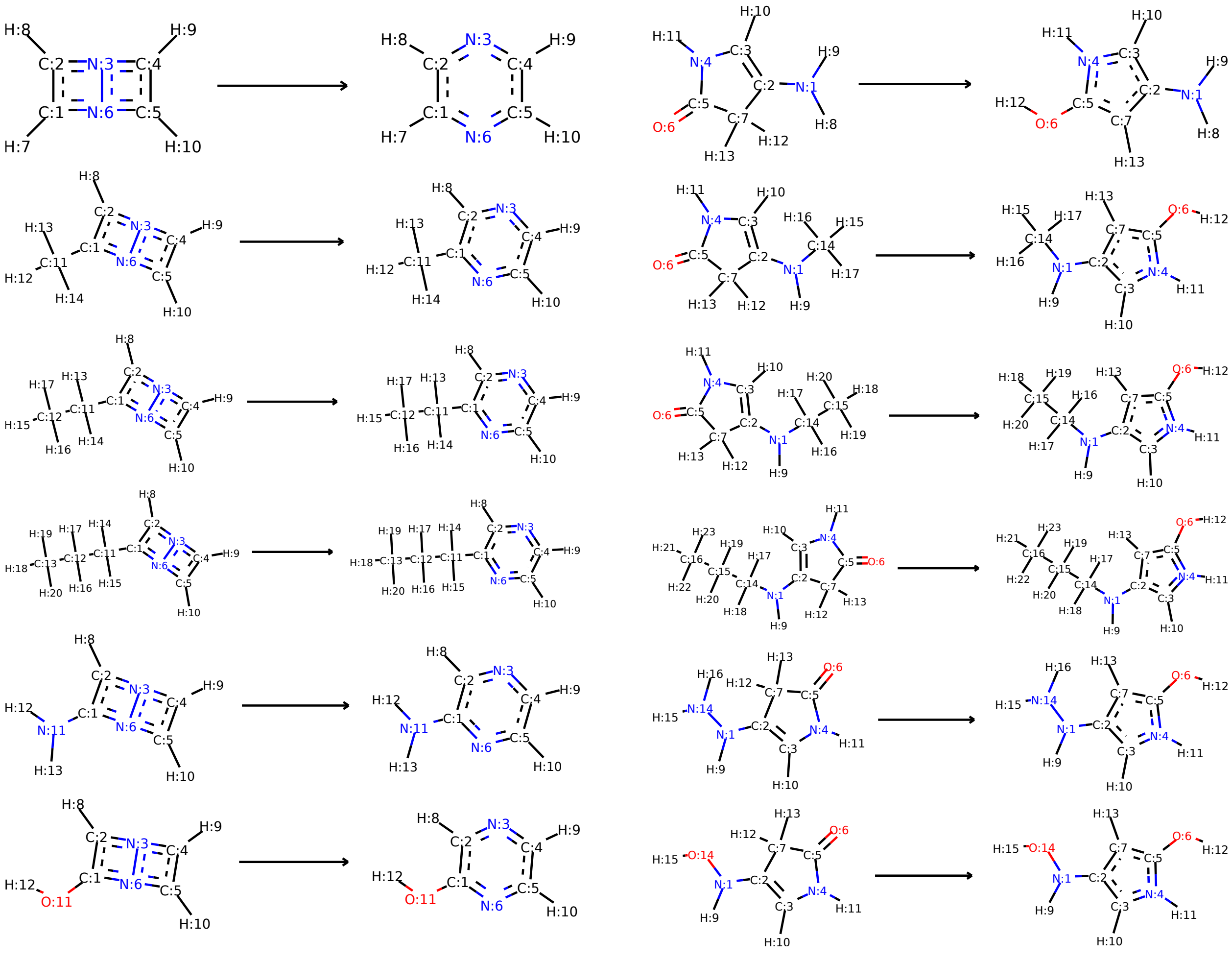


Figure 5.13. Substitution of side chains at a location close to the reactive center (hydrogen 7) for an example reaction (left) and at a location far from the reactive center (hydrogen 8) for a different example reaction. The topmost reactions are the original reactions and the following reactions involve different substitutions of the hydrogen atoms.

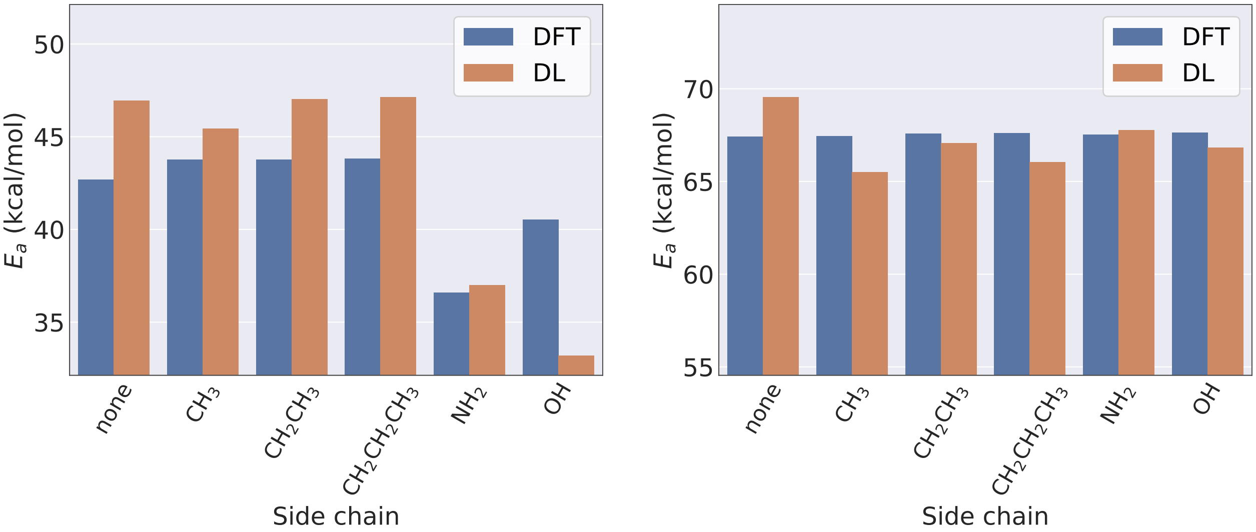


Figure 5.14. Change in activation energy due to the side chain substitutions illustrated in Figure 5.13. The left plot corresponds to the left reactions in Figure 5.13 and the right plot corresponds to the right reactions in Figure 5.13. The "true" activation energies for the substituted reactions were calculated using DFT (ωB97X-D3/def2-TZVP) and are compared to the deep learning (DL) predictions. Note that the ordinate in both plots is scaled such that both plots have the same spacing and that its range goes from 15 kcal mol^{-1} below the maximum barrier to 5 kcal mol^{-1} above the maximum barrier, but does not start at zero.

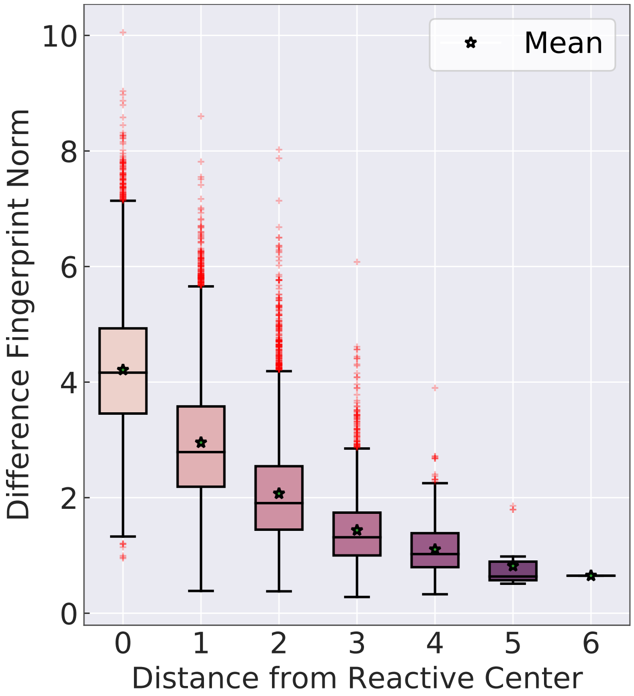


Figure 5.15. Investigation of how the difference fingerprint norm, i.e., contribution to the activation energy prediction, changes as atoms move farther from the reactive center. The reactive center is defined as those atoms that undergo bond changes in a reaction.

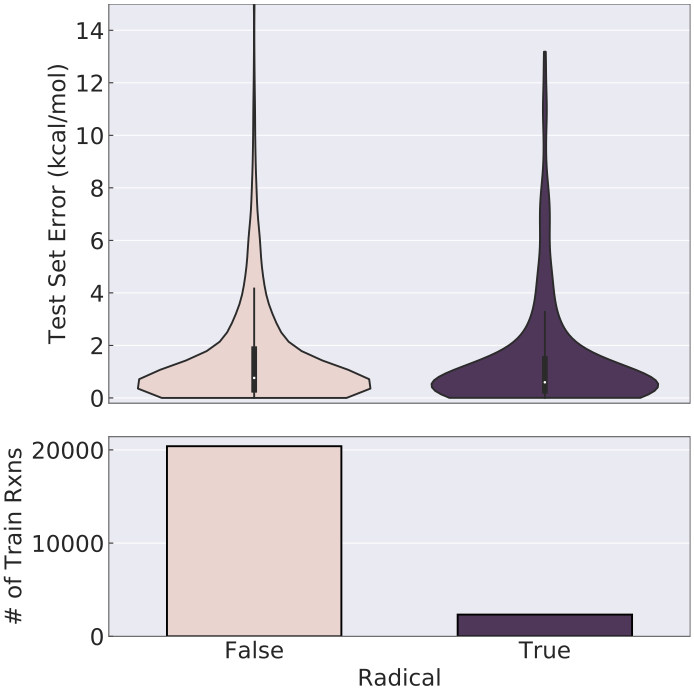


Figure 5.16. Comparison of test set error distributions for non-radical and radical reactions.

atoms farther from the reactive center, although some distant atoms may influence the reactivity via electronic or steric effects. Figure 5.15 shows that the contribution, measured by the norm of the difference fingerprint, does indeed decrease for atoms that are farther from the reactive center.

5.3.4 Radical reactions

The model reported in the previous sections was not trained using the additional radical reaction data. To investigate whether accurate predictions are also possible for reactions containing radicals, we combined the two data sets described in Section 5.2.2 and retrained the model. The results, split by whether or not the reaction contains radicals, are shown in Figure 5.16 for a single fold. Even though the combined data is composed only of 11% radical reactions, there is no significant difference between the performance on reactions that contain radicals and those that do not. Therefore, this model is a suitable candidate for use in automated mechanism generation, where radical reactions are very common.

5.4 Conclusion

With quantitative data becoming more readily available through advances in high-throughput experimentation and more extensive computational resources available for data generation using, for example, quantum chemistry, quantitative predictions of reaction performance based on large training sets are becoming increasingly more feasible. Here, we have demonstrated that activation energies for a diverse set of gas-phase organic chemistry reactions, including radical reactions, can be predicted accurately using a template-free deep learning method. We expect that automated reaction mechanism generation software can strongly benefit from such a model, whether to estimate kinetics or to enable discovery of new reactivity. Further generation of large quantitative data sets will likely result in rapid development of novel machine learning algorithms suitable for predicting such quantities.

5.5 References

(1) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540.

(2) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442–452.

(3) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.

(4) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113.

(5) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(6) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(7) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.

(8) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using a Combined Linguistic Model and Hyper-Graph Exploration Strategy. **2019**, arXiv: 1910.08036.

(9) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis Across Pharma Chemical Space. *Chem. Commun.* **2019**, 12152–12155.

(10) Chen, B.; Shen, T.; Jaakkola, T. S.; Barzilay, R. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis. **2019**, arXiv: 1910.09688.

(11) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D., The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information In *The Future of the History of Chemical Information*; American Chemical Society: Washington, DC, 2014; Chapter 8, pp 127–148.

(12) Mayfield, J.; Lowe, D.; Sayle, R., Pistachio: Search and Faceting of Large Reaction Databases In *American Chemical Society National Meeting*, Washington, DC, 2017.

(13) Evans, M. G.; Polanyi, M. Intertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(14) Sumathi, R.; Carstensen, H. H.; Green, W. H. Reaction Rate Prediction via Group Additivity Part 1: H Abstraction from Alkanes by H and CH_3. *J. Phys. Chem. A* **2001**, *105*, 6910–6925.

(15) Sumathi, R.; Carstensen, H. H.; Green, W. H. Reaction Rate Prediction via Group Additivity, Part 2: H-Abstraction from Alkenes, Alkynes, Alcohols, Aldehydes, and Acids by H Atoms. *J. Phys. Chem. A* **2001**, *105*, 8969–8984.

(16) Saeys, M.; Reyniers, M. F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. Ab Initio Group Contribution Method for Activation Energies for Radical Additions. *AIChE J.* **2004**, *50*, 426–444.

(17) Sabbe, M. K.; Vandeputte, A. G.; Reyniers, M. F.; Waroquier, M.; Marin, G. B. Modeling the Influence of Resonance Stabilization on the Kinetics of Hydrogen Abstractions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1278–1298.

(18) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(19) Gastegger, M.; Marquetand, P. High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015**, *11*, 2187–2198.

(20) Allison, T. C. Application of an Artificial Neural Network to the Prediction of OH Radical Reaction Rate Constants for Evaluating Global Warming Potential. *J. Phys. Chem. B* **2016**, *120*, 1854–1863.

(21) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. - Eur. J.* **2018**, *24*, 12354–12358.

(22) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catal. Lett.* **2019**, *149*, 2347–2354.

(23) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(24) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**.

(25) Collins, K. D.; Gensch, T.; Glorius, F. Contemporary Screening Approaches to Reaction Discovery and Development. *Nat. Chem.* **2014**, *6*, 859–871.

(26) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(27) Grambow, C.; Swanson, K.; Yang, K.; Hirschfeld, L.; Jin, W. chemprop-reaction: Release version 1.1.0, version 1.1.0, Zenodo. https://doi.org/10.5281/zenodo.3712473, 2020.

(28) Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. **2015**, arXiv: 1509.09292.

(29) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

(30) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**, arXiv: 1704.01212.

(31) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed 08/06/2018).

(32) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **2013**, *53*, 2812–2819.

(33) Osório, N.; Vilaça, P.; Rocha, M., A Critical Evaluation of Automatic Atom Mapping Algorithms and Tools In *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Fdez-Riverola, F., Mohamad, M. S., Rocha, M., De Paz, J. F., Pinto, T., Eds.; Springer International Publishing: 2017, pp 257–264.

(34) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic Mapping of Atoms Across Both Simple and Complex Chemical Reactions. *Nat. Commun.* **2019**, *10*.

(35) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(36) St. John, P.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. Prediction of Homolytic Bond Dissociation Enthalpies for Organic Molecules at near Chemical Accuracy with Sub-Second Computational Cost. **2019**, DOI: 10.26434/chemrxiv.10052048.

(37) Grambow, C. A. Reactants, Products, and Transition States of Radical Reactions, version 1.0.0, Zenodo. https://doi.org/10.5281/zenodo.3731553, 2020.

(38) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(39) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(40) Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Chapter 6

Recommendations for future work

Through a combination of quantum chemistry and machine learning, this thesis has developed improved methods to discover new reactions and compute thermochemical and kinetic parameters that may be important in high-temperature radical-driven mechanisms, atmospheric chemistry systems, and potentially even organic synthesis. It has shown that quantum chemical automated transition state finding algorithms are systematic and thorough at finding new reactions, especially those with low barriers, and can be used in a massively parallel approach to generate tens of thousands of reactions spanning a wide range of chemical space not subject to human bias or intuition. With machine learning research permeating the chemical sciences and a plethora of novel methods being developed, this thesis has capitalized on the quickly growing research by implementing new methods for thermochemical and kinetic parameter estimation. We have shown that thermochemistry can be estimated to near chemical accuracy with machine learning and that automated bond additivity corrections can significantly improve the estimates provided by cheap quantum chemistry methods. We demonstrated that a deep learning model for the estimation of activation energies is suitable to significantly scale up the discovery of chemical reactions. Nonetheless, several limitations of the methods and models remain, especially in the machine learning area. Overcoming these challenges in the future will be crucial for more accurate and more pervasive reaction discovery, and chemical parameter estimation.

6.1 Enhancing bond additivity corrections

Chapter 3 demonstrated that relatively cheap quantum chemistry methods, for example, coupled cluster with small basis sets or density functional theory methods, can provide very accurate enthalpies of formation using simple bond additivity correction (BAC) schemes. However, in order to be useful for the entire community, BACs must be derived for many different combinations of quantum chemistry methods and basis sets. Deriving an exhaustive set of BACs is nearly impossible due to the sheer number of available methods and basis sets, but selecting the most promising methods with the most promising basis sets will enable a comprehensive comparison that can be used to determine which method and basis set are most suitable given a desired accuracy and computational cost. Additionally, the BACs in Chapter 3 should be enhanced by including a better torsional treatment when computing the partition functions from quantum chemistry calculations, as enthalpy of formation can also be influenced by internal rotations.[1] Specifically, one-dimensional hindered rotor treatments are a popular, automatable method for calculating torsional partition functions.[2,3] Over time, a large database of corrections should be accumulated in the Arkane code available in the RMG software.[4]

Error-canceling balanced reactions provide an alternative to BACs for calculating accurate enthalpies of formation.[5,6] Isodesmic reactions, which conserve the types of bonds in a reaction, are an example of such reactions and can also be derived in an automated fashion.[7] Implementing such a scheme and comparing it to BACs will also be important for the thermochemistry and automated mechanism generation community so that the best method can be chosen in any given situation.

While simple parametric BAC models have been successful for many systems and methods, more complicated methods, especially neural networks, may be a promising alternative for deriving BACs that would not require defining empirical functional forms and may be able to overcome some of the limitations described in Section 3.6.

6.2 Improving reaction discovery and kinetics estimation

6.2.1 Generating multimolecular reactions

Chapters 2 and 4 showed that automated transition state finding algorithms can be used to discover extensive sets of reactions in *unimolecular* systems. While the products of these reactions may be multimolecular, true *multimolecular* reactions with more than one reactant and more than one product have not been generated in large numbers. In theory, the current methods can be directly applied to systems containing two or more reactions, but it can be reasonably assumed that reaction discovery would not be efficient due to the significant increase in the number of degrees of freedom as a result of the many different orientations two or more molecules can have toward each other. For example, if the reacting sites on two molecules are pointed away from each other, an automated transition state finding method will have to undergo additional work to rotate the molecules into a suitable orientation lest the transition state optimization fail. Additionally, it stands to reason that the discovered reactions could be biased more significantly toward high-barrier processes because there exist factorially more combinations of how to pair reactants together. Future work should address these issues by judiciously choosing reactant combinations and developing heuristics for orienting reactants. The work by Dewyer and Zimmerman can serve as a suitable starting point for this problem.[8,9]

6.2.2 Estimating kinetics

Chapter 5 provides the first important step toward temperature-dependent kinetics estimation by predicting activation energies, but complete rate estimation requires at least the prediction of the Arrhenius pre-exponential factor in addition to the activation energy. An alternative approach could be to directly predict the molecular partition functions of the reactants and transition state of the reaction.[10] A flexible and general approach for estimating rates would be desirable in automated mechanism generation because the existing approach is based on template-specific hierarchical trees of rate estimation rules, which have to be constructed manually, although automated approaches for decision tree generation are also being explored.[11]

6.3 Addressing limitations of machine learning in chemistry

The encouraging results of Chapters 3 and 5 indicate that machine learning methods are well-suited for automated mechanism generation and automated reaction discovery. Nonetheless, there exist many limitations that must be overcome for its continued success and to enable its widespread application.

6.3.1 Building interpretable models

In chemistry, an abundance of physical models with clearly-defined, explanatory equations exist, which naturally enables chemically meaningful interpretation of the model predictions. In deep learning, the opaque decision-making process of neural networks leads to a lack of explanatory models that are not useful for guiding human reasoning.[12,13] In fact, it is precisely due to the existing physical models that the improvement afforded by deep learning in chemistry is not as significant as in computer science where physical models are not common.[12] While this may currently limit the usefulness of novel methods, there exists significant potential for progress in this area.[14] Several papers that try to combine the benefits of neural networks with physical models are starting to be published. Pfau et al. developed an approach to solve the Schrödinger equation without the need for a finite basis set by using a neural network ansatz for the wave function which obeys the required Fermi-Dirac statistics.[15] Their model can outperform coupled cluster theory in strongly-correlated systems but has to be retrained for each new molecule, which severely limits its applicability. Schütt et al. overcome this issue by training a model to directly predict the Hamiltonian matrix from which chemical properties can be derived that are defined as quantum mechanical operators on the wave function.[16] Sinitskiy and Pande also overcome the limitation by training a deep neural network to compute electron densities in addition to energies.[17]

Instead of integrating chemical domain knowledge with deep neural networks, several approaches to enable interpretability revolve around analyzing the importance of hidden units[18] or adding layers that quantify the contributions of each atom.[19–21]

Ultimately, additional research is required in both of the areas mentioned above to satisfy the needs of chemists and chemical engineers attempting to use such models for novel discovery.

6.3.2 Overcoming data scarcity

While further advances in method development are necessary for the continued success of machine learning in chemistry, additional data sources and superior ways of dealing with small data are just as important. One of the most significant issues is the significant difficulty associated with obtaining large quantities of chemical data. The reasons for this are that obtaining data is costly,[13,22] currently available data is severely limited compared to data sets in computer science,[12] and it can even be difficult to validate model results because success can often only be proved using experiments.[23] Furthermore, training data are often inherently biased and can lead to non-generalizable models.[13] For example, machine learning models may learn irrelevant patterns as demonstrated in the yield prediction study by Ahneman et al., which showed that their models were learning the hidden structure inherent in their data set.[24–27] If training data are not carefully scrutinized, it can also be very easy to train a model that learns a nonsensical prediction. For example, a seemingly accurate model of protein affinity predictions can be obtained due to redundancies in the training data even if the data involve affinities toward many different proteins but does not provide the protein target as input to the model.[28,29] It should be noted that the authors of the studies in Refs. [28] and [29] do not mention this inconsistency in their papers.

Luckily, significant progress toward improving data limitations is being made and does not always require more data. Transfer learning, as described in Chapter 3, is one way of dealing with data scarcity, but other approaches exist. A promising direction that should be explored in further research uses unsupervised techniques, especially autoencoders and generative models, to obtain a transferable continuous molecular representation that could be used as input for subsequent supervised problems or could be used to directly sample new molecules or reactions with desirable properties.[30] There exist at least 166 billion organic molecules with up to 17 heavy atoms.[31] Training an unsupervised model to learn a molecular representation on significant fractions of those molecules could potentially lead to much more generalizable models. While such large models do not yet seem to be common in chemistry, with the largest ones using only one million molecules,[32] models with hundreds of millions of parameters trained on billions of words are now being published routinely in the area of natural language processing.[33,34] Extending models trained on relatively small molecules to work with larger molecules may be possible using hierarchical modeling, which has been demonstrated successfully for polymers.[35]

Although improving transferability will certainly lead to better models, knowing when to trust the model predictions is extremely important both for deciding whether to use a prediction in an actual application and for choosing which new data points to train on.[14] Uncertainty quantification of neural network predictions is becoming more pervasive, but the estimated uncertainty and the true error are often not strongly correlated, thereby leaving room for significant improvement.[36,37]

6.4 References

(1) Li, Y.-P.; Bell, A. T.; Head-Gordon, M. Thermodynamics of Anharmonic Systems: Uncoupled Mode Approximations for Molecules. *J. Chem. Theory Comput.* **2016**, *12*, 2861–2870.

(2) Pitzer, K. S.; Gwinn, W. D. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation: I. Rigid Frame with Attached Tops. *J. Chem. Phys.* **1942**, *10*, 428–440.

(3) McQuarrie, D. A., *Statistical Mechanics*; University Science Books: 2000.

(4) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(5) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular Orbital Theory of the Electronic Structure of Organic Compounds. V. Molecular Theory of Bond Separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.

(6) Wheeler, S. E.; Houk, K. N.; Schleyer, P. V.; Allen, W. D. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *J. Am. Chem. Soc.* **2009**, *131*, 2547–2560.

(7) Buerger, P.; Akroyd, J.; Mosbach, S.; Kraft, M. A Systematic Method to Estimate and Validate Enthalpies of Formation Using Error-Cancelling Balanced Reactions. *Combust. Flame* **2018**, *187*, 105–121.

(8) Dewyer, A. L.; Zimmerman, P. M. Finding Reaction Mechanisms, Intuitive or Otherwise. *Org. Biomol. Chem.* **2017**, *15*, 501–504.

(9) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *8*, e1354.

(10) Desgranges, C.; Delhommelle, J. A New Approach for the Prediction of Partition Functions Using Machine Learning Techniques. *J. Chem. Phys.* **2018**, *149*, 044118.

(11) Johnson, M. S.; Green, W. H., A Machine Learning Based Algorithm for Rate Estimation In *AIChE National Meeting*, Orlando, FL, 2019.

(12) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307.

(13) Rodrigues Jr, J. F.; Florea, L.; de Oliveira, M. C. F.; Diamond, D.; Oliveira Jr, O. N. A Survey on Big Data and Machine Learning for Chemistry. **2019**, arXiv: 1904.10370.

(14) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem. Int. Ed.*, DOI: 10.1002/anie.201909989.

(15) Pfau, D.; Spencer, J. S.; Matthews, A. G. d. G.; Foulkes, W. M. C. Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks. **2019**, arXiv: 1909.02487.

(16) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(17) Sinitskiy, A. V.; Pande, V. S. Physical Machine Learning Outperforms “Human Learning” in Quantum Chemistry. **2019**, arXiv: 1908.00971.

(18) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R., Quantum-Chemical Insights from Interpretable Atomistic Neural Networks In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K.-R., Eds.; Springer, Cham: 2019; Vol. 11700, pp 311–330.

(19) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(20) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, DOI: 10.1021/acs.jmedchem.9b00959.

(21) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule Attention Transformer. **2020**, arXiv: 2002.08264.

(22) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.

(23) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(24) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C−N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186–190.

(25) Chuang, K. V.; Keiser, M. J. Comment on “Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning”. *Science* **2018**, *362*, eaat8603.

(26) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on “Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning”. *Science* **2018**, *362*, eaat8763.

(27) Cova, T. F.; Pais, A. A. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* **2019**, *7*, 809.

(28) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(29) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(30) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—A Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849.

(31) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(32) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction: Next-Gen QSAR Models with MolPMoFiT. **2019**, DOI: 10.26434/chemrxiv.9978743.

(33) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. **2018**, arXiv: 1810.04805.

(34) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. **2019**, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

(35) Jin, W.; Barzilay, R.; Jaakkola, T. Hierarchical Generation of Molecular Graphs using Structural Motifs. **2020**, arXiv: 2002.03230.

(36) Ryu, S.; Kwon, Y.; Kim, W. Y. A Bayesian Graph Convolutional Network for Reliable Prediction of Molecular Properties with Uncertainty Quantification. *Chem. Sci.* **2019**, *10*, 8438–8446.

(37) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning Based Molecular Property Prediction. *J. Chem. Inf. Model* **2020**, DOI: 10.1021/acs.jcim.9b00975.